DOCUMENT RESUME

ED 413 733 FL 024 455

AUTHOR	Guerrero, Michael D.
TITLE	A Critical Analysis of the Validity of the Four Skills Exam.
PUB DATE	1994-12-00
NOTE	194p.; Doctoral Dissertation, University of New Mexico, Albuquerque.
PUB TYPE	Dissertations/Theses - Doctoral Dissertations (041) -- Reports - Evaluative (142)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	Bilingual Education; Comparative Analysis; Construct Validity; Elementary Secondary Education; Grammar; Language Proficiency; Language Skills; *Language Tests; Listening Comprehension; Native Language Instruction; *Native Speakers; Reading Comprehension; *Second Languages; *Spanish; Spelling; *Test Reliability; *Test Validity; Vocabulary
IDENTIFIERS	*Four Skills Exam

ABSTRACT

A study evaluated the overall evaluative validity of the Four Skills Exam, a Spanish language proficiency test designed to ensure that bilingual education teachers in New Mexico can meet Spanish language demands in the bilingual education classroom. The test's construct validity was limited for several reasons. In designing a test capturing real-life language demands, developers did not operationalize the targeted demands effectively. The two objectively scored parts of the test yielded unacceptable reliability coefficients. Internal consistency of the subjectively scored parts was spuriously high due to a halo effect and absence of explicit scoring benchmarks. A moderately high correlation between aural and reading parts was found. One analysis found that examinees who grew up speaking Spanish and spoke it currently in the home performed no better than those lacking these experiences. Content identified for the test was not fully embedded, what was incorporated was being used for the wrong grade levels, and it was skewed toward vocabulary, spelling, and grammar. It is concluded that making valid inferences concerning the language abilities of the examinees based on test scores is difficult, the social consequences of using pass-fail scores are undesirable, and the test is not adequately filling its intended purpose. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 733

A CRITICAL ANALYSIS
OF THE VALIDITY OF THE FOUR SKILLS EXAM

By

Michael D. Guerrero

Bachelor of Arts, Spanish Language & Bilingual Education
Eastern Michigan University, 1980

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

Michael D.
Guerrero

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy in Education

The University of New Mexico
Albuquerque, New Mexico

December 1994

FL024455

BEST COPY AVAILABLE

© 1994, Michael D. Guerrero

ACKNOWLEDGEMENTS

Consuelo, my wife, is without a doubt the key person responsible for helping me bring this dissertation to closure. She supported me unconditionally, day in and day out, and through the most difficult times. I must also pay tribute to Andrea, my daughter, for her keen awareness of what I was doing all those hours in the study room and for her understanding and support. As a family we completed "Daddy's book".

To Dr. John W. Oller, Jr., the Chair of my committee, I owe an incalculable amount of gratitude and appreciation. Dr. Oller gave unselfishly of his time, expertise, wisdom, patience and praise, and I needed all he was able to give. Dr. Oller is a great man, and it has truly been an honor to fall under his tutelage.

I also owe a tremendous amount of gratitude to the remaining three members of the committee. I must thank Dr. Ortíz for supporting me unconditionally over the years. To Dr. Chris Nelson for all the time he spent with me making sense of the data and helping me maintain a healthy attitude. To Dr. Luisa Durán, I owe much appreciation for her contributions throughout this process. I also owe many thanks to Thomasina Hannum, one of the original authors of the Four Skills Exam, and to Archie Griñe and Jun Barrack in the Testing Division for all their support and assistance.

The support I received from my family and Committee was bolstered by support from many other professors, colleagues and friends. I must recognize the significant role played by Dr. Paul Martínez and all my colleagues at EAC-West. I must also give full recognition to Dr. Kathy Escamilla, Dr. Leonard Baca and Dr. Jim Bransford for their support and contributions to this work.

A CRITICAL ANALYSIS
OF THE VALIDITY OF THE FOUR SKILLS EXAM

By

Michael D. Guerrero

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy in Education

The University of New Mexico
Albuquerque, New Mexico

December 1994

BEST COPY AVAILABLE

A Critical Analysis of the Validity of the Four Skills Exam

Michael D. Guerrero

Bachelor of Arts, Spanish Language & Bilingual Education
Eastern Michigan University, 1980
Doctor of Philosophy in Education, Educational Linguistics
University of New Mexico, 1994

Using the framework advanced by Messick (1989), this dissertation set forth an overall evaluative judgment of the validity of the Four Skills Exam, a Spanish language proficiency test designed to ensure that bilingual education teachers in New Mexico are able to meet common Spanish language demands characterizing a bilingual education classroom. A variety of evidence was generated which revealed certain weaknesses in this high-stakes test and its applications.

The construct validity of the test fell short for several reasons. First, the original test development team sought to develop a test which captured real life language demands, but fell short of effectively operationalizing the targeted demands. The internal consistency of each subtest was statistically examined (n= 217). The two objectively scored parts of the test yielded less than acceptable reliability coefficients. The internal consistency of the subjectively scored parts of the test were arguably spuriously high owing to a halo effect and the absence of explicit scoring benchmarks. A moderately high correlation between the aural and reading parts of the exam was also found. Moreover, one of several MANOVAs was conducted which indicated that those examinees which grew up speaking Spanish and reported speaking Spanish presently in the home performed no better than those examinees

lacking these experiences.

Regarding the content relevance of the exam, the test development team also failed to fully embed the content they had originally identified. Furthermore, the content of the test was aimed at the early elementary grades (K-4), but upon completion of the test the state department of education endorsed the use of the test for grades K-8; it is presently being used K-12. The content coverage of the test was found to be redundant since it is heavily skewed towards the measurement of vocabulary, spelling and grammar.

The validity of the test is also viewed in light of the roles of the institutes of higher education. The statistical analyses performed indicated that overall the examinees are not well prepared, especially in the areas of Spanish literacy.

Given the above discrepancies, it proves difficult to make valid inferences regarding the language abilities of the examinees based on their test scores. The meaning of the test scores is blurred due to the instrument's psychometric shortcomings and the use of the instrument for purposes for which it was not designed.

The social consequences of using the pass and fail scores generated by the examinees taking the Four Skills Exam are undesirable. The intent driving the development and adoption of the test was to protect the rights of Spanish speaking children in need of instruction in their native language. Given the evidence set forth, there is little reason to believe that the exam is adequately fulfilling this intended social function. There is also some evidence which indicates that those examinees who are Hispanic surnamed, grew up speaking Spanish, still speak Spanish at home, and are native to NM experience difficulty in passing the Four Skills Exam. These individuals represent the pool of potential role models for the students in need of Spanish language instruction, but are being screened out by a measure of questionable validity.

In sum, the Four Skills Exam appears to have outlived its usefulness owing to its original design and to the inappropriate use of the test by policy making entities. Future research of a qualitative nature is sorely needed.

TABLE OF CONTENTS

List of Figures xi

List of Tables xii

Chapter 1: Introduction 1

Chapter 2: Review of the Literature 9

  Theoretical foundations of language proficiency 10

  Language test development 20

  Bilingual teacher target language skills 26

  Non-English proficiency criteria in New Mexico 27

  The development of the Four Skills Exam 31

  Description of the Four Skills Exam 37

  Theoretical orientation of the Four Skills Exam 44

  Content relevance and coverage of the Four Skills Exam 48

  Technical standards of the Four Skills Exam 51

  Spanish language proficiency testing in the Southwest 63

  Test validity in social context 71

  Summary: Review of the literature 72

Chapter 3: Methods 78

  Subjects 78

  Materials 81

  Procedure 82

  Analyses 88

Chapter 4: Reliability Analyses and Results 90

Common parts 91

Uncommon parts: Form A 99

Uncommon parts: Form B 108

Summary 115

Chapter 5: Correlations and Equivalency of Forms 119

Correlations: Form A 120

Correlations: Form B 135

Equivalency of Forms 147

Summary 149

Chapter 6: Additional Aspects of Validity 150

Overall performance of examinees 150

The role of institutes of higher education 152

Formal Spanish language training 154

Spanish language background 157

Geographic location 158

Ethnicity and test performance 160

Summary 162

Chapter 7: Conclusions and Future Research 165

References 176

LIST OF FIGURES

Figure 1 Components of Communicative Language Ability in Communicative Language use 11
Figure 2 A Hierarchical Semiotic Model of Language Proficiency 19

LIST OF TABLES

Table 1 Test Sites across New Mexico 79

Table 2 Geographic Location Where Examinees Grew Up 79

Table 3 Grade Level Distribution of Examinees 81

Table 4 Reliability: Part 1 (Aural) Listening Comprehension 92

Table 5 Reliability: Part 1 (Aural) Informal Words 93

Table 6 Reliability: Part 1 (Aural) Formal Equivalents 94

Table 7 Reliability: Part 2 (Oral) Passage One 96

Table 8 Reliability: Part 4 (Composition) 97

Table 9 Reliability: Part 1 (Aural) Dictation 100

Table 10 Reliability: Part 2 (Oral) Passage Two 101

Table 11 Reliability: Part 2 (Oral) Passage Three 102

Table 12 Reliability: Part 3 (Reading) Orthography: Accents 103

Table 13 Reliability: Part 3 (Reading) Orthography: Spelling 105

Table 14 Reliability: Part 3 (Reading) Identifying Concepts 106

Table 15 Reliability: Part 3 (Reading) Words in Context 107

Table 16 Reliability: Part 1 (Aural) Dictation 109

Table 17 Reliability: Part 2 (Oral) Passage Two 110

Table 18 Reliability: Part 2 (Oral) Passage Three 110

Table 19 Reliability: Part 3 (Reading) Orthography: Accents 111

Table 20 Reliability: Part 3 (Reading) Orthography: Spelling 112

Table 21 Reliability: Part 3 (Reading) Identifying Concepts 113

Table 22 Reliability: Part 3 (Reading) Words in Context 114

Table 23 Summary of Reliability Across Test Forms 116

Table 24 Correlations Between Aural and Oral Common Parts and Subtests and Common Parts 122

Table 25 Correlations Between Reading Comprehension Subtests, Test Parts, and Common Parts 124

Table 26 Correlations Between Test Parts and Common Parts Variable . . . 125

Table 27 Correlations Between Aural Subtests and All Remaining Test Variables 126

Table 28 Correlations Between Oral Passages and All Remaining Test Variables 129

Table 29 Correlations Between Reading Subtests and Test Parts and Total Score Variable 131

Table 30 Correlations Between Test Parts and Composite Test Variables 133

Table 31 Correlations Between Composite Variables 135

Table 32 Correlations Between Aural Subtests and Remaining Test Variables 136

Table 33 Correlations Between Oral Passages and Remaining Test Variables 140

Table 34 Correlations Between Reading Subtests and Other Test Variables 141

Table 35 Correlations Between Parts and Composite Test Variables 143
Table 36 Correlations Between Composite Variables and Total 146
Table 37 ANOVA Source Table for Scores Across Test Forms 148
Table 38 ANOVA Source Table for Scores on Uncommon Parts Variable 149
Table 39 Pass/Fail Percentages for Examinees 151
Table 40 Univariate Source Table for Site Variable 153
Table 41 Univariate Source Table for Formal Study of Spanish 155
Table 42 Univariate Source Table for Spanish Language Background 158
Table 43 Univariate Source Table For Residence Variable 159
Table 44 ANOVA Source Table for Ethnicity Variable 161

CHAPTER 1

INTRODUCTION

The Four Skills Exam is a Spanish language proficiency test widely used in New Mexico to determine whether or not a prospective bilingual education teacher has acceptable Spanish language proficiency (Valdés, 1989) to meet the demands of a bilingual education classroom setting. The goal of this study was to make a series of sound judgments regarding the unified validity (Messick, 1989) of this test. Such an overall evaluative judgment required the analysis of a variety of evidence which might add to or detract from the validity of the psychometric instrument. Before detailing the specific methods of this dissertation, it may be useful to get a better picture of the problems it addresses.

Across the state of NM different groups of prospective bilingual education teachers are taking the Four Skills Exam on designated test dates and at designated test sites. After approximately two and a half hours, the cassettes used for oral parts of the test and all the exam forms are collected by the proctors. The proctors then mail the collected materials to a designated test center. The test materials are shortly made available to the two designated test scorers in the state. The two test scorers then make a determination from the scores as to whether or not each examinee failed all or some portions of the test. Some of the part scores are generated objectively as there is only one correct answer to the items on those parts; other part scores are arrived at subjectively since the scorer must use his or her best judgment guided by selected criteria. Eventually, the results are made known to each examinee by mail or phone. Some examinees pass the test, or some portions of it, while others fail it in its entirety. Depending on the results of the exam, that is the scores given to the examinee by the test scorer, the prospective bilingual education teachers move closer to their desired goal—bilingual endorsement and potential employment.

In effect, the professional destiny of the prospective bilingual teacher in NM depends, in part, on the ability to pass the Four Skills Exam. However, there is more at stake than just the teachers' professional and economic destiny. Teachers meeting this requirement hopefully will go on to serve the educational needs of the student community. More importantly, the professional and economic future of the students served depends in part on the ability of the bilingually endorsed teacher to communicate and deliver instruction in the Spanish language.

Therefore, the Four Skills Exam is a high-stakes test. It affects the lives of prospective bilingual education teachers, the students to be served and eventually the community as a whole when those students enter the work force. Important social consequences hinge on the reliability of the test scores, the appropriateness of the test items on which the test scores are founded (i.e., construct validity), and their interpretation. For all these reasons, it is important to examine the overall validity of the Four Skills Exam, or what Messick (1989) calls its "unified validity".

To gauge the "unified validity" of a test, Messick (1989) posits the following four-faceted question:

> The four inter-related aspects of this question ask what balance of evidence supports the interpretation or meaning of the scores, what evidence undergirds not only score meaning, but also the relevance of the scores to the particular applied purpose and the utility of scores in the applied setting; what rationales make credible the value implications of the score interpretation and any associated implications for action; and what evidence and arguments signify the functional worth of the testing in terms of its intended and unintended consequences. (p. 5)

In order to address Messick's first question—what balance of evidence supports the interpretation or meaning of the scores—one must first consider the instrument's reliability. As Bachman (1990) says:

> If test scores are strongly affected by errors of measurement, they will not be meaningful, and cannot, therefore, provide the basis for valid interpretation or use. A test score that is not reliable, therefore, cannot be valid. (p. 25)

For insight into Messick's second question—what is the relevance of the scores to the particular applied purpose and setting—one must seek answers to several other questions. Bachman (1990) points out that:

> In order for a test score to be a meaningful indicator of a particular individual's ability, we must be sure it measures that ability and very little else. Thus, in examining the meaningfulness of test scores, we are concerned with demonstrating that they are not unduly affected by factors other than the ability being tested. (p. 25)

For example, does a passing score on the oral section of the test mean that the prospective bilingual education teacher can deliver instruction in Spanish? To adequately address this question one must look for supporting evidence related to the construct validity of the instrument, including the content relevance and coverage, and testing formats.

With regard to Messick's third question (i.e., what rationales make credible the value implications of the score interpretation and any associated implications for action), multiple value judgments are made related to score interpretation. For example, a passing score on the written section of the test should mean that the examinee can meet the written demands associated with a bilingual education setting. However, the criteria underlying a `Pass' score reflect the values of the test developers. To what degree are these value judgments defensible or credible? What evidence is there to support these value judgments?

Finally, the fourth question is what evidence and arguments signify the functional worth of the testing in terms of its social consequences? Does the test serve its intended purpose and what is the supporting evidence? What unintended social consequences has the testing process generated and what evidence is there to demonstrate them?

As Messick (1988: 42) says and schematically represents:

> Test validity, as an overall evaluative judgment of the adequacy and appropriateness of inferences and actions based on test scores, thus rests on four bases.... Putting these four bases together, we see that test validity can be represented in terms of two interconnected facets linking the source of the justification-either evidential or consequential-to the function or outcome of the testing, either interpretation or use. This crossing of basis and function provides a unified view of test validity, as portrayed in Fig. 3.1 (Messick, 1980).

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Value Implications	Social Consequences

Fig. 3.1. Facets of test validity.

Problems to be Addressed

Given this framework and the testing enterprise under consideration, the following tasks were undertaken with the purpose of setting forth an overall evaluative judgment of the validity of the Four Skills Exam. Specifically, the following series of theoretical, psychometric and statistical analyses were addressed in this dissertation:

1. Statistical analyses of the reliability of the instrument were conducted using Cronbach's α. These analyses included an item-analysis, tests of internal consistency, and equivalency across the three alternate forms of the test.
2. Analyses of the construct validity of the instrument, including content relevance and coverage and testing formats were conducted. These analyses focused on the theoretical foundation of the instrument, the definitions and operationalization of the constructs purported to be measured by the instrument's test formats, and the content relevance and coverage of the test items. Pearson's *r* was used to examine the correlations between the subtests and parts of the test. In addition, multivariate analyses of variance (MANOVA) were conducted to address two questions:
    - A. Did the test performance of the examinees vary as a function of formal language training?
    - B. Did the test performance of the examinees vary as function of Spanish language background?
3. An examination of the value implications underlying the interpretation of test scores was conducted. This analysis was centered on examining the value implications underlying a pass or fail score in the four skill areas measured by the exam.

4. An examination was undertaken to identify the intended and unintended social consequences of using the exam. A series of MANOVA and ANOVA statistical analyses were conducted to determine the relationship between sociodemographic variables and test performance. Specifically, the following questions were raised:

   A. What was the Pass and Fail rate of the examinees on the different parts of the test and the test as a whole?

   B. Did the performance of the examinees vary as a function of institutional affiliation?

   C. Did the performance of the examinees vary as a function of the region of the state in which they happened to reside?

   D. Did Hispanic surnamed examinees perform differently than the non-Hispanic surnamed examinees?

Summing up, the primary objective of this dissertation was to set forth evidence, following Messick's (1989) framework, which speaks to the validity of the Four Skills Exam in the particular social and political context of New Mexico. The evidence has been garnered by recurring to state of the art knowledge and research related to language proficiency theory, language measurement and bilingual education. Evidence has also been generated by conducting necessary statistical analyses entailing key variables.

Justification of the Study

The need for this study stems from several sources. First, one of the central authors of the Four Skills Exam, Dr. Guadalupe Valdés, recommends that the effective-ness of the test be periodically reexamined (Valdés, 1989). However, the reliability or validity of this test has *not* been reexamined since its adoption by the NM State Board of Education in 1981. Second, while tests like the Four Skills Exam are widely used in the U.S. and the Southwest in general, there is very limited research on the Spanish language proficiency exams used for endorsing bilingual education teachers. Third, and as set forth by the National Commission on Testing and Public Policy (1990), "Rarely is an important test or its use subject to formal, systematic, independent professional scrutiny or audit" (p. 21). Finally, there is evidence of a growing concern among the stakeholders in New Mexico regarding the validity of the Four Skills Exam. In 1993 a Task Force was assembled by the New Mexico Association for Bilingual Education to reexamine the validity of the Four Skills Exam and to make recommendations to enhance the validity of the instrument in question.

Limitations of the Study

The fundamental limitation of this study is the fact that while the test has been used in New Mexico for approximately thirteen years, only test data from 1991-1992 were used in this dissertation. The reason for selecting this time frame is based on the assumption that the examinees would have received similar Spanish language training since the adoption of the Spanish language competencies in 1989 by the NM State Department of Education. All of the available tests from that period were used. A random selection was neither feasible nor desirable given the limited number of tests available. Nonetheless, the sample size ($n = 217$) was adequate to support the analyses required. A second limitation of the study concerns the review of the actual oral and written protocols generated by the examinees. It was beyond the scope of this dissertation to reexamine the scoring of the oral and written parts of the test in order to assess the inter-rater reliability of these two test parts. On the other hand, only two scorers were used to score these two parts of the test and their separate or individual ratings were not available. Nonetheless, ample evidence was secured in order to make judgments about the overall reliability and validity of these two parts of the Four Skills Exam relative to each other and the other part scores available. Lastly, the design did not include interviews or questionnaires with examinees which would surely have enhanced the judgments about the validity of the test.

CHAPTER 2

REVIEW OF THE LITERATURE

This chapter consists of several topics related to the task of moving towards an overall evaluative judgment of the validity of the Four Skills Exam. The first section focuses on theoretically defining language proficiency. The second section concerns the central challenge in language test development, the operationalization of the construct (i.e., language proficiency) to be measured. The third part of this chapter provides insight into the kinds of language skills and abilities bilingual education teachers should have.

The fourth section reviews the development of the Four Skills Exam, including a description of the different parts of the test and what is known about the psychometric and technical properties of the instrument. The fifth part of this chapter offers a comparison between the Four Skills Exam and three other tests used for the same purpose in the Southwest. The final part of this section provides food for thought regarding the sociolinguistic milieu in the U.S.. The chapter concludes with a summary highlighting each of the four areas (i.e., construct validity, content relevance and utility, value implications and social consequences) Messick (1989) considers central to examining the unified validity of a test.

Theoretical Foundations of Language Proficiency

In order to measure language proficiency it is crucial to first have an understanding, a conception, a theoretical model of language proficiency. This model must be consistent with how language is structured and used in actual human behavior. As Oller & Damico (1991) state:

> According to Cronbach (1970), it is necessary to develop theoretical notions, which he called "constructs": theoretical factors posited as organizers or controllers of some aspect of behavior. The "construct(s)" posited must with some determinable reliability be seen in real life performances. (p. 79)

There are two prevailing models of language proficiency which address the issue at hand. Bachman (1990) has developed what he describes as a theoretical framework of communicative language ability, while Oller (1991) posits a hierarchical model of language proficiency based on pragmatic theory. Each model is examined below.

Bachman (1990) describes his model of communicative language ability (CLA) in the following manner:

> The framework of CLA I propose includes three components: language competence, strategic competence, and psychophysiological mechanisms. Language competence comprises, essentially, a set of specific knowledge components that are utilized in communication via language. Strategic competence is the term I will use to characterize the mental capacity for implementing the components of language competence in contextualized communicative language use. Strategic competence thus provides the means for relating language competencies to features of the context of situation in which language use takes place and to the language user's knowledge structures (sociocultural knowledge, `real-world' knowledge). Psychophysiological mechanisms refer to the neurological and psychological processes involved in the actual execution of language as a physical phenomenon (sound, light). (p. 84)

BEST COPY AVAILABLE

Bachman (1990: 85) schematically represents the interaction of these components with the language use context and the language user's knowledge structure as represented in Figure 1.

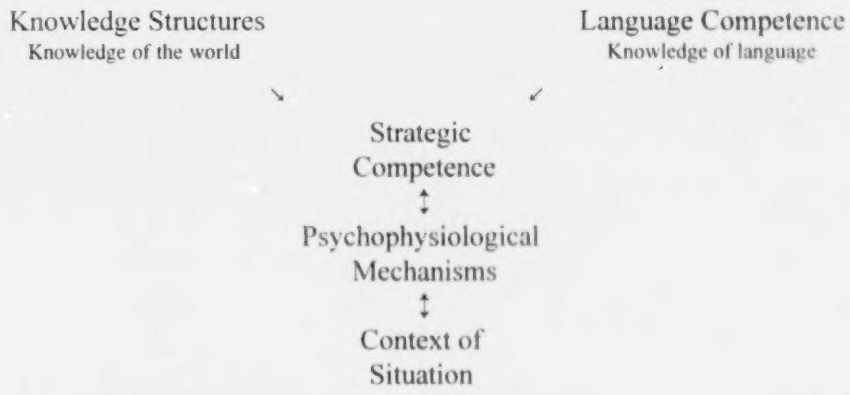
Knowledge Structures
Knowledge of the world
Language Competence
Knowledge of language
Strategic Competence
Psychophysiological Mechanisms
Context of Situation

Figure 1 Components of communicative language ability in communicative language use adapted from Bachman, L. (1990). Fundamental considerations in language testing. Oxford University Press.

The strength of this model lies in the detailed description of the language competence component. Briefly, according to Bachman language competence consists of two general competencies: organizational competence and pragmatic competence. Under the organizational component belong the grammatical and textual competencies. The grammatical competence, in turn, consists of one's competence in vocabulary, morphology, syntax and phonology/graphology. Textual competence subsumes the cohesion and rhetorical competencies.

Under pragmatic competence fall illocutionary competence and sociolinguistic competence. The former competence subsumes a series of language functions: ideational, manipulative, heuristic, and imaginative. It also subsumes one's competence to produce and comprehend direct and indirect speech acts. Sociolinguistic competence consists of four sub-competencies: sensitivity to dialect or language variety, register, and naturalness. The fourth sub-competency consists of one's ability to produce and comprehend language linked to cultural references and figures of speech. In fact, this conception of language competence clearly draws from four salient disciplines of linguistic inquiry: structural linguistics, discourse analysis, the functions of language and sociolinguistics.

The central problem with Bachman's (1990) model concerns the function of the strategic competence component. Bachman describes this component as a `mental' capacity or as "the means for relating language competencies to features of the context of situation in which language use takes place and to the language user's knowledge structures (sociocultural knowledge, `real-world' knowledge)" (p. 84). In effect, the strategic competence component of this model is the central component since it mediates between the remaining four components displayed above.

Briefly, strategic competence consists of three sub-components: an assessment, planning and execution component. Regarding the role of the assessment component, Bachman (1990) states:

> The assessment component enables us to (1) identify the information-including the language variety or dialect- that is needed for realizing a particular communicative goal in a given context; (2) determine what language competencies (native language, second or foreign language) are at our disposal for most effectively bringing that information to bear in achieving the communicative goal; (3) ascertain the abilities and knowledge that are shared by our interlocutor; and (4) following the communication attempts, evaluate the extent to which the communicative goal has been achieved. (p. 100)

The assessment component appears to involve a considerable amount of mental activity (i.e., identifying needed information, selecting the most effective language resources, ascertaining information about the interlocutor, and evaluating the communicative outcome). How is all this accomplished in the absence of experiential knowledge, memory and intelligence? Bachman makes no reference to the need to access the knowledge structures component of his model to assess a given communicative event. He does, however, briefly mention the role of intelligence in relation to strategic competence by saying that it may be inaccurate to identify it with intelligence.

The role of memory is alluded to under the planning component of strategic competence. First, Bachman states that "The planning component retrieves relevant items (grammatical, textual, illocutionary, sociolinguistic) from language competence and formulates a plan whose realization is expected to achieve the communicative goal" (p. 101). The retrieval of such information clearly implies the role of memory, both short and long term. However, Bachman makes no direct reference to the role of memory in this model.

The last subcomponent of strategic competence according to Bachman is the execution component. It entails the use of relevant psychophysiological mechanisms to carry out the plan in the appropriate language modality (i.e., receptive and productive) and channel (i.e., visual and auditory). Bachman (1990) explains:

> In receptive language use, auditory and visual skills are employed, while in productive use the neuromuscular skills (for example, articulatory and digital) are employed. (p. 107)

A brief comment regarding the psychological process of 'visual skills' in this model is warranted. Bachman (1990) characterizes 'visual skills' as the speaker's ability to gain key non-linguistic information from the communicative context. However, he explicitly dismisses non-verbal manifestations of strategic competence from his model, while acknowledging the importance of non-verbal communication. This inconsistency detracts from the credence of Bachman's model. Moreover, the model is incomplete in the absence of some viable characterization of non-verbal communication as it relates to language proficiency.

For the purposes of this dissertation, Bachman's model of communicative language ability offers a conceptual baseline of language competence which may prove useful in examining this particular aspect of the validity of the Four Skills Exam. On the other hand, the core component of the model, strategic competence, is superficial or not fully developed. This is evident as the author implies the role of other human cognitive capacities such as memory, experience, non-verbal communication, and intelligence but fails to articulate their interrelatedness.

Oller (1991) provides a more comprehensive model of language proficiency which is founded on language testing research supporting the existence of a general language factor and pragmatic theory linked to the early thinking of C. S. Peirce. Briefly, and as detailed in Oller and Damico (1991), recent language testing research reveals that diverse language tests are positively and considerably correlated. This finding supports the idea that there is a common factor inherent to diverse language measures. Nonetheless, research also shows that language tests measure specific language abilities which are distinct from this common factor. Consequently, Oller posits the existence of a general factor that is decomposable into various yet integrated components. He uses this empirical language testing research to substantiate the existence of both general and specific factors which must be accounted for in an adequate theory of language proficiency.

In contrast to Bachman (1990), Oller (1991) postulates an intimate relationship between language proficiency and intelligence. This relationship or interdependency between language and thinking (i.e., intelligence) is, as Oller (1991) states, succinctly summarized in a citation from Einstein (1941):

> ...Everything depends on the degree to which words and word combinations correspond to the world of impression.
>
> What is it that brings about such an intimate connection between language and thinking? Is there no thinking without the use of language, namely in concepts and concept combinations for which words need not necessarily come to mind? Has not everyone of us struggled for words although the connection between "things" was already clear?
>
> We might be inclined to attribute to the act of thinking complete independence from language if the individual formed or were able to form his concepts without the verbal guidance of his environment. Yet most likely the mental shape of an individual growing up under such conditions would be very poor. Thus we many conclude that the mental development of the individual and his way of forming concepts depend to a high degree upon language (1941, in Oller 1989b p. 62). (p. 12)

As Oller surmises, the central assumption is thus that language is the core representational or semiotic medium which nurtures thinking, conceptual development or intelligence as an individual experiences the world. This is not meant as a causal, unidirectional relationship but rather as evidence of the close link between language and intelligence.

The central pragmatic theoretical orientation of Oller's model of language proficiency rests on the concept of pragmatic mapping alluded to in the above paragraph. Pragmatic mapping is concerned with explaining how an individual is able to take in raw sensory data related to the individual's world of experience and convert this information into experience, and eventually, into comprehensible text(s) in a natural language.

Oller refers again to the work of Einstein to describe what has been termed a "gulf" between the world of experience and the semiotic representations one is able to generate in order to represent human experiences. In answer to this problem Oller (1991) has conceived of a theoretical model of language proficiency which consists of four distinct yet interrelated semiotic capacities. This model consists of a general semiotic capacity which mediates between the world of experience (sensory motor images) and its interpretation (semiotic representations of the images). One of the central functions of the general semiotic capacity is to govern three subordinate yet interrelated representational capacities: a sensory motor, kinesic and linguistic semiotic capacity.

Considering briefly the concept of the sensory-motor semiotic capacity, Oller maintains that humans generate iconic representations of experiences. People have the ability to symbolically represent a sensory motor activity and to follow the text of the activity. For example, as one is running along a wooded path, the runner adjusts the pace according to twists in the path (i.e., a visual representation), the texture of the surface (i.e., a tactile representation), the sound of his or her own breathing (i.e., an auditory representation), etc.. Each adjustment is based on sensory motor information that the individual is processing in order to execute the act of running on this particular occasion. In short, this capacity underlies the ability to execute the many routine activities (e.g., driving, shaving, changing a diaper, etc.) an individual acquires in order to get along in the world.

It is important to note that sensory-motor representations are for the most part in a continuous state of flux and out of necessity degenerate . That is, the facts of experience which they represent must fade in order to accommodate the steady flow of sensorial input.

On the other hand, it is also this capacity that enables a person to mentally reconstruct past or to create future and imaginary events. Moreover, it is this capacity which contributes to an individual's ability to share experiences with other people.

The kinesic semiotic capacity concerns a person's ability to receptively or productively process gestural representations (i.e., non-verbal communication) which convey conventional meaning. As Oller and Damico (1991) indicate there is much research which demonstrates how gestures are closely coordinated with a sequence of linguistic forms. During any linguistic exchange, a poorly timed frown, wink, protruding tongue, smile, or act of pointing, will surely gain the listener's attention. Similarly, the inappropriate use of gestural representations (e.g., a hug in place of a hand shake) may lead to extensive negotiation of meaning.

While kinesic representations are clearly conventional, their use in human communication is not always entirely requisite. A speaker may be able to achieve a communicative act with or without the aid of gestural representations. Further, the meaning of a smile, clinched hands, or an extended stare, may be open to a broader range of interpretations than the meaning of an utterance.

The linguistic semiotic capacity comprises the third representational capacity central to Oller's model of language proficiency. Oller and Damico (1991) describe this component in the following manner:

> By contrast with the other types, these achieve a higher level of abstraction and greater potential validity. While sensory motor representations are iconic (analogues of what they mean) and kinesic representations are often ambiguous and require leaps of inference, linguistic representations are typically more abstract and potentially more determinate. (p.90)

Linguistic representations possess a quality of permanency unlike any of the other two representational capacities. As Oller states, a linguistic representation is much more likely to convey the same propositional meaning regardless of the passage of time. Most importantly, and again in contrast to the other representational capacities, linguistic representations can be used to express any factual or fictional idea.

The general semiotic capacity is superordinate to the three semiotic capacities described above. The principal function of this capacity is to orchestrate the overall integration of the subordinate representational systems. Generally speaking, and in any given communicative act, all three capacities are needed. On the other hand, at times it is necessary to engage the different semiotic capacities independently of each other but simultaneously. For example, the sensory motor capacity may be fully engaged while performing an act (e.g., driving, jogging, etc.) while carrying on a conversation unrelated to this activity. Again, the general semiotic capacity oversees these operations. The primary evidence supporting the existence of a general semiotic capacity stems from the fact that an individual is able to generate sensory motor representations triggered by another interlocutor's narrative. This explains the ability people have to visualize a scenario triggered by oral and written discourse. Similarly, an individual may respond kinesically to discourse or sensory motor stimulus. For example, the facial expressions (e.g., grimacing, surprise, disbelief, etc.) of an individual are commonly generated by the discourse (e.g., a novel, speech, etc.) or actions (e.g., an accident, close play in a soccer game, etc.) of another person. This "intertranslatability" of semiotic capacities must be governed by a more general capacity.

In Figure 2, Oller (1991: 18) offers the following modular representation of information processing based on the concept of pragmatic mapping discussed thus far.

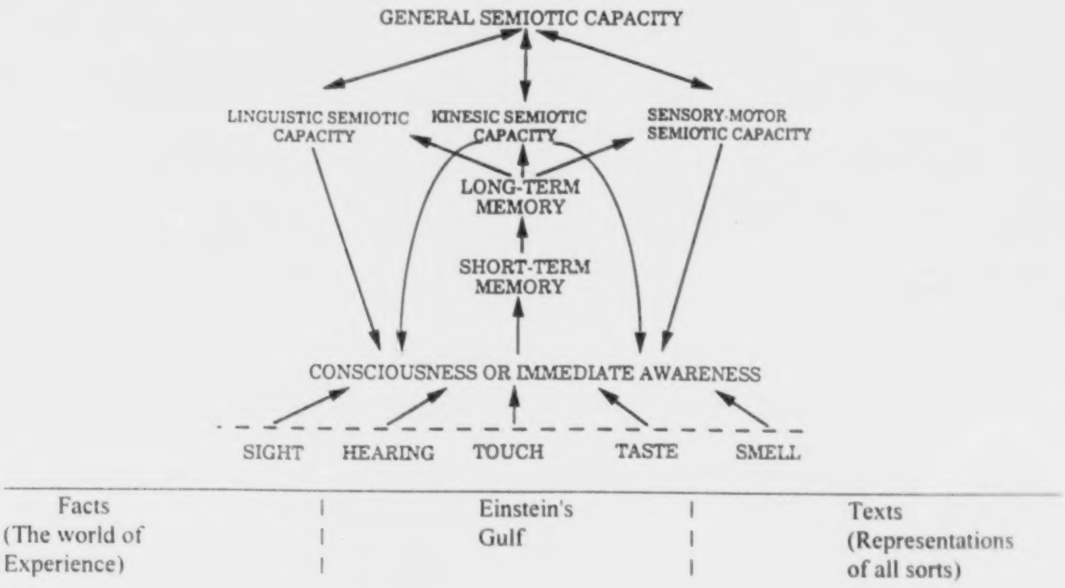
GENERAL SEMIOTIC CAPACITY
LINGUISTIC SEMIOTIC CAPACITY
KINESIC SEMIOTIC CAPACITY
SENSORY-MOTOR SEMIOTIC CAPACITY
LONG-TERM MEMORY
SHORT-TERM MEMORY
CONSCIOUSNESS OR IMMEDIATE AWARENESS
SIGHT
HEARING
TOUCH
TASTE
SMELL
Facts (The world of Experience)
Einstein's Gulf
Texts (Representations of all sorts)

Figure 2 A hierarchical semiotic model of language proficiency adapted from Oller, J.W. Jr. (1991). Language and bilingualism: More tests of tests. Bucknell University Press. Associated University Presses.

Oller (1991) describes the information processing in the following paragraph:

> The focal element in this diagram is consciousness or immediate awareness. The question addressed is how information from the senses is processed via the kinds of grammatical structure that are supplied by the various semiotic systems--linguistic, kinesic, and sensory motor. The idea is that the determination of the meaning of texts is chiefly a matter of relating them via representational capacity with the facts of experience and vice versa. As new texts or representations are processed, they are fed into short-term memory and some of them from there into a longer term memory. Consciousness and memory, together, interact with semiotic systems so as to modify them. Presumably, this is the basis for the acquisition of the conventional aspects of semiotic systems. (p. 17)

In sum, this hierarchical model of language proficiency posited by Oller (1991) clearly offers a complex and abstract yet behaviorally relevant explanation of language

BEST COPY AVAILABLE

ability. The model accounts for the essential components (i.e., intelligence, experience, innate capacities, linguistic and non-linguistic conventions, and memory) which all play a central role in language proficiency. The model is, in this respect, a more coherent whole than Bachman's model reviewed above. The challenge for language testing, however, is to develop a measure which is true to such a theoretical model of language proficiency.

Language Test Development

Briefly, the fundamental task confronted by test developers is minimally threefold. First, a defensible theoretical conception of language proficiency must be identified in order to lay the foundation for the instrument. Again, such a theory must have behavioral relevance. Second, the linguistic demands germane to the targeted contexts of communication must be identified. Third, these linguistic demands must be operationalized in such a way within the test format so that they do not lose their behavioral relevance. Each of these tasks contribute to the construct validity of the test. Oller and Damico (1991) summarize the task as follows:

> To the extent that the theory is on the right track and the assessment procedure is a valid implementation of the theory, results obtained will enable consistent (reliable) and accurate (valid) predictions about actual capabilities and performances of students. This is the same as saying that the "test" (or assessment procedure) will be a valid measure of its construct(s) and that it will enable accurate prediction beyond the testing situation. The acid test of any theory, of any construct that is part of a theory, or of any test or measurement procedure based on the theory is its behavioral relevance (Cronbach, 1970). (p. 80)

In order to make an accurate prediction about one's language proficiency which is based on the individual's test performance, the test must have demonstrated behavioral relevance. Behavioral relevance is in fact synonymous with the concept of 'authenticity' in language testing. Bachman (1990) states:

> One of the main preoccupations in language testing for the past quarter of a century (at least) has been a sincere concern to somehow capture or recreate in language tests the essence of language use, to make our language tests 'authentic'. (p. 300)

The problem which arises is defining what is meant by authentic. Can one test be more authentic than an other? If so, on what criteria can such a judgment be based? Bachman examines two approaches for defining authenticity in language tests, the 'real life' approach and the interactional/ability approach.

Briefly, the 'real life' approach to defining authenticity rests on the degree to which the test performance replicates the corresponding real life language performance. As Bachman (1990) states, "This approach seeks to develop tests that mirror the 'reality' of non-test language use..." (p. 301). In this approach authenticity hinges on the face validity of the test or the match between the testing format and actual non-test context. For example, in the case of bilingual education teachers, a real life measure might entail the creation of a situational context (e.g., preparing a science lesson plan) which ultimately requires the test-taker to read a short science text passage, develop a lesson plan, write out some comprehension questions, and so on.

In contrast, the interactional/ability approach to defining authenticity in language tests focuses on the extent to which language abilities are inherent to the testing situation. Stated differently, the interactional/ability approach is founded on a theoretical framework of language proficiency or the construct(s) (e.g., grammatical competence, kinesic semiotic capacity, etc.) the test is intended to measure.

Bachman (1990) explains that the most salient difference between the two approaches lies not in the differences of testing formats but rather in the linguistic criteria used to score the test taker's performance. The author compares the oral language scoring rubrics of the advanced level of the ACTFL (American Council on the Teaching of Foreign Languages, 1986) with a three part oral language rating scale developed by Bachman and Palmer (1983). The fundamental difference between the two scoring rubrics is evident in that the ACTFL scoring rubric views oral language proficiency as a unitary ability while Bachman and Palmer's view of oral language proficiency is componential in nature consisting of grammatical, pragmatic and sociolinguistic competencies.

Bachman (1990), in closing his discussion on authenticity, suggests a synthesis of the two approaches. The author states:

> The characterization of authenticity is undoubtedly one of the most difficult problems for language testing, as it necessarily involves the consideration of not only the context in which testing takes place, but also of the qualities of the test taker and of the very nature of language ability itself. (p. 330)

Given this brief review of this test quality termed authenticity, authenticity is most appropriately construed in terms of two characteristics: the degree of congruence between the testing context and the social context in which the same non-test language is manifested and the operationalization of the linguistic construct(s) elicited during language testing. For example, and related to contextual authenticity, a testing situation for bilingual education teachers would most likely require the examinee to demonstrate his or her ability to engage in writing activities which characterize the writing demands routinely placed on bilingual educators.

By the same token, the authenticity of the theoretical construct of written expression underlying the instrument would depend on the manner in which the construct has been operationalized. This would preclude testing formats (e.g., multiple choice formats) that do not require the examinee to write extended discourse, for example. Moreover, it would bar limiting the scored criteria to orthography, syntax and vocabulary at the expense of criteria such as coherence, organization and expression.

Shohamy and Reves (1985) outline five factors which reduce `authentic language' to what they and others (for example Spolsky, 1985) term `authentic test language'. The first factor concerns the goal of the interaction underlying a language testing situation. As the authors point out, the purpose of language testing is ultimately to obtain a test score which is based on the quality of targeted linguistic criteria (e.g., intonation, grammar, fluency, etc.), clearly an unnatural communicative purpose in actual communication.

The second factor these authors cite concerns the social relationship which bonds the participants in a testing situation. As Shohamy and Reves state, "We recognize the fact that the tester and test taker would not necessarily be involved in a similar communicative act with one another in real life" (p. 55). The important point is that the social relationship between a tester and test taker is most likely a formal one and one in which both participants may know nothing about the other.

The physical setting where the testing takes place is a third factor which serves to reduce the authenticity of language produced in a test setting. The setting for most language testing is generally a classroom, office, or language laboratory. While interacting in physical academic settings is common place, there are a host of other physical settings (e.g., at home, restaurants, banks, etc.) in which people routinely interact verbally.

The fourth factor discussed by Shohamy and Reves concerns the topic(s) on which the test taker must focus while demonstrating his or language ability. As the authors state, the topic(s), and again the targeted linguistic criteria, are most likely predetermined and imposed by the tester. As the authors state the topics around which real life communication generally takes place are unplanned or determined mutually by the participants.

The final factor cited as a threat to the authenticity of language produced under even the most seemingly authentic language testing situations is a temporal factor. There is generally a time limit imposed on language tests that takes on a different form in actual communication. In contrast, in actual communication, for example, an individual may ask the interlocutor to repeat a statement, provide additional clarification, or to continue the interaction at a later point in time; these possibilities do not generally apply to testing situations.

There are other factors which would clearly reduce the authenticity of even the most sophisticated authentic language testing formats which were not mentioned by Shohamy and Reves (1985) which merit reflection. Specifically, the consequences or outcomes associated with real life communication as opposed to a language testing situation may vary greatly. This is especially true of high-stakes language tests where some type of professional licensure (e.g., a bilingual teaching endorsement) and the potential economic livelihood of the test taker lies in the balance. Rarely does real life communication entail such consequences.

In the final analysis, the goal in measurement is to take a valid theory of the construct under consideration and to design an assessment procedure that is an accurate implementation of the theory (Oller & Damico, 1991). In the case of language proficiency, the construct of interest in this dissertation, even the most authentically seeming testing procedures will inevitably fall short of this elusive goal (Shohamy and Reves, 1985). Consequently, language test development is reduced to aiming for the construction of 'authentic-seeming tasks' (Spolsky, 1985) which can only elicit authentic 'test language' (Shohamy & Reves, 1985). Nonetheless, gauging the authenticity of a test can still yield evidence which is essential for piecing together a global assessment of the unified validity of the measurement instrument.

At this juncture two principal issues have been addressed. The first issue concerns a theoretical conception of what language proficiency might consist of and how its components interact in receptive or productive language use. The model developed by Oller (1991) provides a defensible and coherent framework. The inevitability of not being able to construct a language test with absolute construct relevance has also been discussed under the concept of 'authenticity'. This fallibility is not limited to the task of language measurement, however. Nonetheless, thinking related to authenticity provides a two dimensional basis for making a judgment regarding the unitary validity of a test: the authenticity of the test context and theoretical construct being measured. It is appropriate to now examine what a bilingual education teacher is expected to be able to do in a non-English language.

Bilingual Teacher Target Language Skills

Experts in the field of bilingual education have long expressed, in general detail, the non-English language skills a bilingual education ought to have under control in order to teach in a bilingual setting (Carrasquillo & Segan, 1982; Clark, 1990; Garcia, 1992; Garza

& Barnes, 1989; Trueba, 1989). The Center for Applied Linguistics (1974), nearly two decades ago, described the language proficiency of prospective bilingual/bicultural education teachers in the following two statements:

> The teacher should demonstrate the ability to:
>
> 1. Communicate effectively, both in speaking and understanding, in the languages and within the cultures of both the home and school. The ability will include adequate control of pronunciation, grammar, vocabulary and regional, stylistic, and nonverbal variants appropriate to the communication context.
>
> 2. Carry out instruction in all areas of the curriculum using a standard variety of both languages. (p. 3)

Gaarder (1977) succinctly summarized his position on bilingual education teachers' language proficiency in the following two statements:

> a. They must be native speakers of the other language or have acquired equivalent competence as a prerequisite to entering a training program....
>
> b. They must be literate-able to read and write-in Spanish at least as well as average American school teachers can do these in English. (p. 84)

More recently, the National Association for Bilingual Education (NABE) outlined a series of national standards for the preparation of bilingual/multicultural teachers (1992) Standard 4 addresses the issue of language proficiency and NABE's position is essentially summarized in the following statement:

> Effective bilingual/multicultural teachers have a command of English and a non-English language that allows them to conduct classes in either language with ease and confidence, regardless of level of instruction. This includes using appropriate and varied language at high levels of accuracy and fluency. Bilingual/multicultural teachers understand and accept dialectic differences in students and their families. Further, these teachers have the ability to serve as translators and interpreters for their students and their families. (p. 19)

From each description it is evident that bilingual education teachers, in the opinion of experts in the field, should be held to relatively high standards with regard to their non-English language proficiency ability. These teachers are expected to meet such varied non-English language criteria as to communicate effectively verbally, non-verbally, in reading and writing in varied social and instructional contexts. In effect, bilingual education teachers are expected to be, at the very least, near native speakers, readers and writers (where written forms of the language are extant) of a non-English language. In essence, bilingual education teachers should be as proficient in the non-English language as their monolingual English speaking teacher peers are in English. Nonetheless, these characterizations of the kinds of language skills bilingual education teachers should have in a non-English language do not govern the related language policy in New Mexico.

Non-English Proficiency Criteria in New Mexico

In 1987 the New Mexico State Board of Education adopted the following native language competencies which prospective bilingual education teachers must demonstrate in order to receive a bilingual endorsement to teach in grades K through 8. It is important to mention that these competencies were generated by a Bilingual Education Task Force and the competencies did not actually become effective until July 1, 1989. The language competencies are as follows:

1. Communicate effectively orally and in writing (where written form exists and is allowed*) in the native language. The bilingual teacher:

    a. demonstrates excellent skills of pronunciation and grammar.

    b. utilizes vocabulary appropriate to a broad range of functions, topics, and genres of speech.

c. demonstrates competency as a participant in ordinary social situations in which the Native language is spoken.

*d. responds adequately to written material by exercising the processes of comparing, contrasting, categorizing, summarizing, inferring, analyzing, synthesizing, hypothesizing and evaluating.

*e. reads with comprehension a broad range of literary forms (folk, technical, classic, etc.).

*f. writes sentences, paragraphs, essays, utilizing standard language mechanics which express original thought, communicate complete and well-organized ideas, and accomplish a full set of written functions.

*g. demonstrates at least a minimum eighth grade level of proficiency in the native language in oral and written language skills where the written form exists and is allowed.

2. Carries out instruction in content areas of the curriculum using a standard variety of the Native language.

The above language proficiency competencies are similar to those recommended by the experts in the field described previously (e.g., Center for Applied Linguistics, etc.). Item (2.) of the language competencies mandated by the New Mexico State Board of Education, the ability to teach in the non-English language across the curriculum, is essentially identical to one of the linguistic abilities identified by the Center for Applied Linguistics in 1974 and the National Association for Bilingual Education in 1992.

Subtle differences can be found between these two sets of linguistic criteria regarding pronunciation, grammar and vocabulary. The New Mexico competencies establish a higher standard in these areas as opposed to the Center for Applied Linguistics. The former panel of experts use the modifier `adequate' as opposed to the latter panel's modifier `excellent' with regard to pronunciation and grammar.

Items (b.) and (c.) are similar to the language abilities identified by the Center for Applied Linguistics. In both cases the bilingual education teacher is characterized as needing an ample vocabulary and to be able to communicate competently in situations outside the school setting. It could also be argued that items (d.), (e.) and (f.) of New Mexico's criteria would satisfy the position taken by Gaarder (1977). That is, that bilingual teachers should be as proficient in the non-English language as monolingual teachers are in English in the areas of reading and writing.

At this point, there is substantive overlap between the criteria advanced by experts and the competencies endorsed by the New Mexico State Department of Education. In both cases, the consensus seems to be that bilingual education teachers should be native or near native speakers, readers and writers of the non-English language.

On the other hand, confusion ensues when one examines item (g.) of the New Mexico criteria. Bilingual education teacher non-English proficiency is abruptly equated with that of an "eighth grade level of proficiency". With the exception of (c.), competencies (a.) through (f.) are not congruent with competency (g.). That is, it seems unlikely that an individual with an eighth grade level of proficiency in any language could demonstrate excellent grammatical skills, have a broad vocabulary, and read and write at the level of competency outlined in items (d.), (e.) and (f.). Similarly, it is unlikely that an eighth grade level of proficiency would suffice to carry out instruction in content areas of the curriculum.

This situation creates at least one serious problem for bilingual education in New Mexico. Which linguistic baseline criteria, the native-like or the eighth grade level, should be used to guide the development of a language proficiency test to uphold the language competencies established by the New Mexico State Board of Education? This issue will be addressed in the following section.

It is important to indicate that in 1989 it became the official responsibility of the state institutes of higher education in New Mexico to assist prospective bilingual teachers in developing these competencies. The colleges and universities in New Mexico which offer the necessary program of studies for a bilingual endorsement most generally embed these competencies within the required courses. For example, competency (2.), regarding delivering instruction in content areas, could be addressed by course work related to bilingual methods or curriculum development. Similarly, competencies related to the development of reading and writing skills could be addressed by requiring the prospective bilingual education teacher to submit written assignments in the non-English language or through course work in the foreign language department.

At a different level, it is the responsibility of the Professional Standards Commission within the New Mexico State Department of Education to ensure that the program of studies leading up to a bilingual endorsement offered by institutes of higher education address the native language competencies outlined above (Scargall, personal communication, 1993). The criteria used by the Professional Standards Commission to determine whether or not the language competencies are readily addressed does not appear to be explicit. Once the Professional Standards Commission approves the program of studies, the State Board of Education must then approve the course of studies.

The critical point here is that there are at least three groups of professional entities (i.e., state college and university faculty, members of the Professional Standards Commission and the State Board of Education) responsible for assisting the prospective bilingual teacher in meeting these language competencies. This fact is important since the roles these three groups play may influence the unified validity of the Four Skills Exam. In this case, the Professional Standards Commission and the State Board of Education are responsible for ensuring that prospective bilingual education teachers have an adequate opportunity to develop their non-English language skills. University faculty have the responsibility of creating these opportunities. As illustrated, the unified validity of the test also hinges on the actions of responsible parties external to the test developers.

The Development of the Four Skills Exam

The need for a standardized Spanish language proficiency test for prospective bilingual educators in the state of New Mexico appears to have its origins in 1978. Valdés (1989), the sole published source of information on the Four Skills Exam, indicates the need for a test of this nature grew out of the recognition that bilingual education teachers in the state possessed low levels of Spanish language proficiency. Valdés (1989) uses the following excerpt from the Albuquerque Journal to set the stage for her article:

> BILINGUAL TEACHING EFFORTS UNDER FIRE
>
> Santa Fe (AP)- None of 136 teachers and aides in bilingual programs in New Mexico's schools who were tested could pass a Spanish reading and writing exam at the fourth grade level, the director of bilingual education for the state Department of Education said.
>
> Henry Pascual concluded that colleges of education are spending a lot of federal money turning out Spanish-English bilingual teachers who don't know much Spanish. (3 October 1978) (p. 207)

Valdés (1989) goes on to describe the relatively low level of Spanish language proficiency typifying bilingual education teachers, graduates of state teacher-training institutions, holding a state bilingual endorsement. Native New Mexican Chicano teachers, while able to speak and understand Spanish, felt uneasy about teaching in Spanish. Many of these teachers could not comprehend written texts at the second or third grade level; many could not write in the language. Anglo teachers had even poorer Spanish language skills; they could, however, read well.

In order to remedy the situation, a committee consisting of key stake-holders from around the state was convened by the state bilingual education director. In addition, and as Valdés (1989) states:

> A special test-development subcommittee was formed, which included the linguist, a statistician, a bilingual educator and a Spanish professor. It was agreed that initial work would involve extensive research to identify what bilingual teachers needed to be able to do with language in order work successfully with monolingual Spanish-speaking children and their parents. (p. 217)

The development of what was going to become the Four Skills Exam began in the fall of 1978 and ended in 1981. The first step in the design of the instrument, as stated above, was to identify the Spanish language skills a bilingual education teacher needs. As Valdés (1989) indicates, research by the committee yielded the following Spanish language functions practicing bilingual teachers should be able to fulfill:

> (1) communicate with young children and their parents; (2) use the language to carry out instruction in the classroom; (3) read and comprehend Spanish language text materials used in bilingual programs; (4) write in Spanish with enough accuracy to teach the writing system to young learners and to be able to write letters and notes to parents. (p. 217-218).

In turn, descriptions of how each of the four functions might be manifested was generated by the committee. The descriptions took the form of a survey that was then mailed to fifty experts in the field of bilingual education around the state. A more refined description of the kinds of Spanish language skills bilingual teachers should be able to demonstrate was generated through the expert opinions of those surveyed. Valdés (1989) summarizes the skills as follows:

> Individuals surveyed agreed that teachers could communicate with young children and their parents if they were able to understand child language and both rural and standard varieties of Spanish, and if they could explain normal school requirements and activities (for example, school-yard policies, pull out programs) to persons not familiar with those concepts. Teachers were described as able to teach in Spanish if they could present material in this language easily and comfortably, without undue pauses and hesitations and without revealing large vocabulary gaps. They were able to read in Spanish if they could draw meaning from texts normally used in third grade classrooms. These texts included math, social studies, and reading materials. Finally, individuals interviewed agreed that in order to teach the Spanish writing system to young children, teachers should be able to spell correctly, use the written accent correctly, proofread material, and find mistakes in children's writing. Moreover, a teacher should be able to write notes home to parents containing few orthographical errors. (p. 218)

This was essentially the core description of Spanish language proficiency which would guide the construction of the test. It was an attempt by the committee to capture the real life language skills a bilingual education teacher should have. As expressed earlier in this section, the next task entailed the development of the test items and testing formats which would hopefully elicit test language that bore some resemblance to the identified real life language skills.

Valdés (1989) states that test construction entailed the development of test items and sample tests which were piloted over a two year period. Again, experts in the field of bilingual education were called upon. This time their task was to judge the relationship of the test items to real life bilingual teacher Spanish language demands. Similarly, practicing bilingual teachers, student teachers and Spanish language majors were given the opportunity to take pilot versions of the test and to comment on the test. Valdés also states that over this two year period item analyses were conducted by the team's statistician in order to modify the examination, a point more fully addressed below.

Once pilot testing was complete, the committee then tackled the task of establishing performance standards. That is, the test development team needed to establish cut-off scores to differentiate between those examinees that were deemed proficient and those that were not. According to Valdés (1989), the test development team used the following strategy to set performance standards:

> The procedure followed in setting such standards involved three steps: (1) tabulating the scores of incumbent teachers who were known to be members of populations II and III (fully proficient teachers); (2) tabulating the scores of incumbent teachers who were known to be members of populations I and IV (limited proficient teachers); (3) submitting the final version of the examination to a group of judges who were asked to take the test themselves, examine the scores made by proficient and limited proficient teachers, and make recommendations about cutoff scores for each test subsection. (p. 221)

Valdés goes on to state that this information was used in establishing the standards and cut-off scores. However, Valdés (1989) also states, "It was felt that in order to ensure statewide support for the examination, a sufficiently large enough number of persons needed to succeed in passing the examination" (p. 221). The author states that a compromise was reached by the committee on this matter. Examinees would be allowed to retake only those sections of the exam which they did not pass instead of having to retake and pass the exam in its entirety.

The development of the Four Skills Exam spanned a three year period, 1978 to 1981. As Valdés states, the development of the test was not conducted in the absence of public controversy. Some stakeholders felt that the test was necessary in order to protect the rights of the Spanish speaking students in need of Spanish language instruction; others, primarily incumbent teachers who would have to pass the test, felt threatened by the forthcoming requirement.

Valdés (1989) also states, in 1981 the test was officially endorsed by the New Mexico State Superintendent of Education. Those examinees passing all sections of the test would hence be considered proficient and able to meet the Spanish language demands commonly encountered in a bilingual classroom setting typical to New Mexico in grades K through 8. The Testing Division at the University of New Mexico was given the responsibility to supervise the administration of the exam twice a year at the state's five institutions of higher education. The author also states that the correction of the examination is conducted by a team of raters under the supervision of one of the co-authors of the test. The cost to examinees is $40.00 for the entire test or $10.00 for each sub-section of the test.

Valdés (1989) concludes her documentation of the development of the Four Skills exam by commenting on the acceptance the test has been met with by the various New Mexican stake-holders. She states:

> ...Five years after it was first developed, it is firmly established as the test that bilingual educators must pass in order to be certified. It is seen as an examination that is fair and actually relevant to teachers as they prepare to work with Spanish monolingual children....
>
> The process by which the test was developed contributed directly to its acceptance. An important first step was that individuals in key positions were able to bring together members of different factions and to provide them with information about the nature of language testing. Equally important was the provision to these groups of a concrete plan for test development that was firmly based on an investigation of actual language use. (p. 224)

While Valdés ends her article on an upbeat note she has overlooked two critical facts which have clear implications for the validity of the Four Skills Exam. First, on various occasions Valdés states that Spanish language functions are relevant to "young learners". In addition, the author also indicates that the reading texts used in the test were selected from third grade classrooms. The point is that the Four Skills Exam was designed to ensure that prospective bilingual education teachers teaching young children, perhaps between Kindergarten and third grade, would be able to meet the routine Spanish language functions associated with these grade levels. The problem is that upon completion of the development of the test in 1981, the New Mexico State Department of Education approved the test to endorse teachers in grades K through eight, a purpose for which the test was not designed. In fact, the test is actually used to endorse bilingual education teachers at all grade levels, kindergarten through twelfth grade (M. J. Habermann, personal communication, 1994).

The same situation appears to have occurred in 1987 when the New Mexico State Board of Education approved the non-English language competencies all bilingual education teachers in New Mexico should be able to demonstrate. Recall that competency (g.) requires the prospective bilingual education teacher to demonstrate at least an eighth grade level of proficiency in the non-English language. More importantly, the remaining competencies all seem compatible with the abilities that come to mind when one considers native or near native language proficiency. The point is that the Four Skills Exam was never designed to uphold these competencies, but it is currently used to do so.

In this context, but through no fault of the test developers, the validity of the Four Skills Exam is clearly weakened. The test development team appears to have proceeded in an appropriate manner in attempting to pinpoint the linguistic demands prospective bilingual education teachers routinely confront in a bilingual education classroom in New Mexico. This portion of test development, while critical to developing a valid and behaviorally relevant test, is only part of the psychometric challenge. The linguistic demands, or language functions in this case, must be operationalized in such a manner that the test tasks are at least seemingly authentic and elicit authentic test language.

Description of the Four Skills Exam

This section contains a concise description of the different parts of the Four Skills Exam. Following the description, a discussion regarding the theoretical foundation of the test and the content relevance and coverage of the test is provided. The section concludes with a summary consisting of findings which bear on the validity of the Four Skills Exam.

There are three equivalent forms (Form A, B, and C) of the Four Skills Exam. Hannum (1993b), one of the members of the original test development team and also one of two official scorers of the exam, recently developed a description of the test for distribution to the examinees prior to taking the test. The test, as its name suggests, is designed to measure four language skill areas: aural, oral, reading and composition. The instrument consists of four separate parts two of which consist of subtests. The oral and composition parts of the test have only one section. It is a timed test and takes approximately two and one-half hours to administer. Each part of the test is described below.

Part 1 (Aural) is tape-mediated and designed to be administered in a language laboratory (Valdés, 1989). Thirty minutes are given over to this portion of the exam. This section of the test consists of four separate subtests: Listening Comprehension, Dictation, Informal Words and Formal Equivalents. The Listening Comprehension subtest consists of twenty multiple choice items. Hannum (1993b) describes this measure as follows:

> This sub-section measures your ability to understand natural conversation and narratives, and to extract subtle meaning. It includes conversations and short narratives presented at normal conversational speed, and represents a variety of interactions. Each situation is heard once. [The people heard on the tape are all native speakers of Spanish; two are from Mexico and the others are New Mexicans.] You will listen to the conversations/narratives and answer multiple choice questions. The questions and answer choices are heard once.

Each examinee receives a test booklet in which further directions are given and in which the examinee is to mark or write his or her answers. In this case, the examinee listens to a conversation or narrative, followed by questions and answer choices provided in Spanish. The examinee must then mark his or her answer on the provided answer sheet consisting of twenty items and the options a,b,c, and d.

The Dictation subtest also consists of twenty items of a fill in the blank nature. The examinee listens to the tape recorded sentences and must write in the word or words which are missing from the written sentence in their test booklet. Such an item might take on the following form:

1. Los jugadores ________________________ muchísimo.

Given the following tape recorded input, *"Los jugadores han peleado muchísimo"*, the examinee would need to supply the verb phrase *"han peleado"*. Each sentence is read and heard only once and the written word(s) must also be spelled correctly in order for the answer to be correct. Hannum (1993b) indicates that the missing words are aimed at orthographic and syntactic elements which traditionally present difficulties in writing for Spanish speakers.

The Informal Words subtest consists of ten items intended to test the examinee's mastery of informal New Mexican Spanish (Hannum, 1993b) which entail archaic forms and lexical borrowings from English. The examinee listens to a sentence which is followed by four words. The examinee must select the word which is the informal or regional equivalent of an underlined word which appears in a sentence in the test booklet. For example the examinee might see and hear, "*Este sartén era de mi abuela*" which is then followed by four aurally presented options such as *(a.) cazuela, (b.) puela, (c.) olla,* and *(d.) tazón.* The examinee must select the option which best represents the informal equivalent of the formal and underlined word.

The Formal Equivalents subtest also consists of ten items. The examinee hears a sentence containing a regional or informal variant of New Mexican Spanish; the informal variant is then repeated at the end of the sentence. The examinee must write and spell correctly the formal equivalent of the informal variant. The examinee might hear, "*Esta puela era de mi abuela*" (*Puela*) to which the examinee should produce in writing "*sartén*".

Altogether, the aural section of the exam consists of four tape-mediated sub-sections with a total of sixty test items. The examinee must answer forty-eight of the sixty items correctly, exactly 80% of the items, in order to pass this section of the exam. It is important to note that the Listening Comprehension, Informal Words and Formal Equivalents subtests are identical across all three forms of the test. The Dictation subtest is, however, different yet supposedly equivalent across the three forms of the test. Part 2 (Oral) is also tape-mediated and requires the examinee to produce and record three brief oral speech samples on three designated topics. Examinees are provided with three written situational descriptions written in English which are intended to guide their oral speech samples. At least one of the oral tasks requires the examinee to use a register appropriate for a child, while the two remaining tasks are adult oriented. Consider the following as an example of the type of situational format used in this section of the test:

> There is going to be an important Parent Advisory Council meeting for the bilingual program at your school. You need to inform and invite all your students' parents. One Spanish speaking parent is dropping off his child so you approach the parent. Explain to the parent that there will be a PAC meeting, an important issue will be discussed requiring parent input, and that transportation and child care can be arranged. Add any additional information you feel might be relevant to the topic.

It is important to note that the first oral task is identical across all three forms of the test. The second oral task is essentially identical across the three forms as well. That is, the language function (i.e., explaining) and the intended interlocutor (i.e., a monolingual Spanish speaking child) on the second oral task are identical across the three forms. They differ primarily in the area of vocabulary. The third oral passage is also similar across the three forms of the test in terms of the designated interlocutor, language function, and vocabulary.

The oral section of the test requires fifteen minutes to administer. The examinee is given a specified amount of time to prepare for each oral task; for one task the examinee is given two minutes to prepare and for the remaining tasks the examinee is given four minutes to prepare for each task. Similarly, the examinee must speak for a specified amount of time on each task; the first task requires the examinee to speak for at least one minute while the two remaining tasks each require a two minute oral protocol. Collectively, the examinee is expected to produce at least five minutes of oral discourse. However, the examinee only needs to pass two of the three passages to pass this part of the test.

Part 3 (Reading) consists of four multiple choice sub-sections. The first sub-section, Orthography: Accents, is designed to measure the examinee's knowledge of orthography, specifically the written accent. In the test booklet the examinee sees a sentence written in Spanish. Some of the words in the sentence require a written accent and in some cases the target words do not. Each sentence is followed by the target words and the examinee must select the option with the correct use of the written accent. For example, a sentence might read:

Este proceso es larguisimo pero si vale la pena.

_____ *1.* *a. larguisimo*
b. larguísimo
c. larguisímo

_____ *2.* *a. si*
b. sí

There are five or six key sentences followed by multiple choice answers. There are a total of twenty items on this sub-section.

The second sub-section, Orthography: Spelling, is also focused on spelling correction and consists of twenty items or sentences. Each sentence contains one spelling error which the examinee must identify and spell correctly in the space provided. There are, however, no errors in the use of the written accent. Such an item might look like the following:

______________ *1. El dicionario contiene miles de palabras.*

Hannum (1993b) indicates that the sentences contain errors that prove problematic for beginning writers.

BEST COPY AVAILABLE

The third sub-section, Reading: Identifying Concepts, consists of ten items which are based on readers, math, social studies, and science texts, including teacher's editions of texts written in Spanish at a third and fourth grade level. Hannum (1993b) describes this subtest as follows:

> ...There are 4 passages to read and identify the principal idea. There are 4 word problems in math to read and identify the process or operation necessary to solve each one. In addition, there are 2 longer passages to read. Each of these longer passages is followed by 4 statements which must be put in the order in which the idea/information appears in the original passage.

The first four test items require the examinee to read a paragraph and to then identify the main idea from among three options (a,b, and c). The next four test items require the examinee to determine which mathematical process (e.g., adding, subtraction, etc.) is required in order to solve a problem presented in sentence form. Each math process is represented by an option (a, b, c, and d). Such an item might take the following form:

5. Hay 60 segundos en un minuto. ¿Cuántos segundos hay en cinco minutos?

____ para saber la respuesta.

The last two test items, as noted above, require the examinee to read two passages approximately 150 words long. Following each passage are four statements which must be placed in the order in which the idea was originally expressed in the text. For each passage, each of the four statements must be in the correct order in order to get the test item correct.

In the final reading subtest, Reading: Understanding Words in Context, the examinee must read a short text in which a word, words or short phrase is missing. There are between five and six passages taken from elementary school science texts on this portion of the test. While the test format appears to be of a cloze nature, the missing items are not systematically deleted as every nth word. Moreover, the examinee is given four options from which to choose.

The third section of the exam is designed to measure the examinee's reading skills and consists of two subtests entailing orthographic skills (i.e., the use of the written accent and spelling correction) and two reading comprehension subtests based on actual Spanish language educational texts used in New Mexican bilingual programs at the third and fourth grade level. It is important to note that this part of the exam is different yet equivalent across the three forms of the test. As in the first part of this exam, this section consists of a total of sixty items and the examinee must answer 48 of the items correctly in order to pass this section of the test. It is important to note that the examinee is given 90 minutes to complete the third and fourth sections of the exam.

The fourth and final section of the Four Skills Exam, Composition, consists of a 150 to 200 word composition (i.e., letter) the examinee must write to parents on one of two predetermined topics written in Spanish. The first topic is essentially the same across all three forms of the test. The second topic again is more alike than different across the three forms of the test; that is, two forms of the test require the examinee to write a letter to the same audience explaining a problem the student is having. The functions entailed in the composition across all three forms are quite uniform.

In sum, the Four Skills Exam is a timed Spanish language proficiency exam designed for prospective bilingual education teachers in New Mexico consisting, as its name suggests, of four parts: aural, oral, reading and composition. There are three equivalent forms of the test, Forms A,B, and C. The test employs a tape-mediated format in the aural and oral sections of the exam. The aural and reading sections are objectively scored. That is, both sections utilize a discrete point format for which there is only one correct answer. The oral and composition sections employ a subjective format for which there are target criteria (e.g., fluency, vocabulary, etc.) to be rated on a five point scale.

Theoretical Orientation of the Four Skills Exam

There is no clear statement regarding the theoretical model of language proficiency which guided the development of the Four Skills Exam. However, Valdés (1989) makes reference to the "functions" (p. 217) of classroom Spanish and the "real-life demands" (p. 220) of the bilingual education teacher. As previously noted, Bachman (1990) defines the "real life" approach to language testing as one that aims to capture the "reality of non-test language use" (p. 301). From the above references made by Valdés (1989) and given the definition provided by Bachman (1990), one can infer that the intent of the test development of the Four Skills Exam was to develop a test that was behaviorally relevant, a test that consisted of authentic seeming tasks (Spolsky, 1985) and elicited authentic test language (Shohamy & Reves, 1985).

While the intent of the test development team may have been to develop a behaviorally relevant test, the task of moving from a theoretical model of language proficiency to its operationalization, as stated previously in this chapter, is one of the greatest challenges facing language test developers. The semiotic model of language proficiency recently advanced by Oller (1991), and previously reviewed, is clearly behaviorally relevant. Moreover, it can serve as a valid model for judging the behavioral relevance (i.e., construct validity) of the Four Skills Exam since the instrument was purportedly founded on real life language use.

In the case of the aural and reading portions of the test, the semiotic linguistic capacity of the examinee is tapped principally at the lexical level. Three of the aural subtests (i.e., Dictation, Informal Words, and Formal Equivalents) elicit primarily vocabulary and spelling knowledge. The two reading subtests with the most test items (i.e., Orthography: Accents and Orthography: Spelling) are also focused on discrete lexical items. It is also important to point out that there is no cohesive and authentic discursive link between the different test items on these subtests. For example, none of the above aural measures are based on a cohesive piece of discourse.

This limited focus on discrete vocabulary items also has the undesirable effect of limiting the integration of language skills. For example, none of the aural subtests are linked to speaking, reading or writing in a behaviorally relevant manner. The Dictation subtest, for example, requires the examinee to write in a missing word, a listening comprehension task which seems remote for bilingual education teachers. Perhaps a more authentic (tape-mediated) task would entail the dictation of a story generated by a student. Such a test format would have the desirable effect of transcending the lexical level within the linguistic semiotic capacity. Similarly, both the aural comprehension and writing skills of the examinee would be called upon interactively.

Much the same case can be made regarding the reading subtests. It seems unlikely that a bilingual education teacher would have to read a short passage and then select a missing word from among three or four options as in the Words in Context subtest. It seems much more likely that a bilingual education teacher would have to read and comprehend an educationally related text and then orally present or discuss the information with the students. or perhaps produce some written comprehension questions based on the reading. Again, the Words in Context subtest does not require the examinee to process language in a behaviorally relevant manner. The semiotic linguistic capacity is engaged primarily at the lexical level and no other language skills such as listening, speaking or writing are activated.

With regard to the process of intertranslatability across semiotic capacities, both the aural and reading parts of the test are centered in the domain of the linguistic semiotic capacity. Neither the kinesic or sensory semiotic capacities are engaged beyond the visual input of the print on the test booklet. The use of pictures, maps, tables, diagrams, video, etc. are totally absent from the Four Skills Exam.

Somewhat in contrast to the aural and reading parts, the oral and composition portions of the exam engage the linguistic semiotic capacity at the text or discourse level. The oral part of the exam which is judged primarily on the linguistic features of fluency, vocabulary and appropriateness does however transcend the lexical level. That is, the examinee must generate brief samples of discourse which require control over syntax, cohesion, function, register, etc..

As in the case of the aural and reading parts of this test, there is no natural engagement of the other three language abilities on the oral part of the test. It is entirely feasible to design an oral measure entailing aural comprehension or reading ability. The oral passages elicited by the test format all reflect the use of oral language in the absence of another interlocutor. A comment is also merited regarding the temporal factor involved in this portion of the test.

That is, one is not always afforded the luxury of time to prepare for engaging in oral discourse. In short, the spontaneous use of oral language is not elicited given the design of the test.

As noted previously, the process of intertranslatability in not readily activated in the Four Skills Exam since there is no sensorial input other than the test print and the aural tape-mediated input. An oral measure could easily accommodate visual input such as pictures, maps, diagrams, etc..

As concerns the composition, a slightly different picture emerges. First, this task requires language processing within the domain of the semiotic linguistic capacity, though at the level of discourse or an extended written text. In addition, there is the integration of reading since the composition prompts are written in Spanish and must be read by the examinee prior to generating the composition. On the other hand, the use of visual stimuli other than the printed prompts is absent. Written products could also be elicited through the use of pictures, diagrams, maps, and tape-mediated input, for example.

In fact, from a more comprehensive theoretical perspective (e.g., Oller, 1991), the Four Skills Exam appears to be founded on a restricted model of language proficiency. This fact critically questions the construct validity or behavioral relevance of the test. Overall, the test development team of Four Skills Exam do not appear to have created the "real life" language tasks they set out to. In other words, the test development team failed to adequately operationalize the language constructs of interest. In general, the language tasks are not authentic and consequently the test language falls short of being authentic test language (i.e., do not match likely school-based language requirements).

Content Relevance and Coverage of the Four Skills Exam

In order to make a fair judgment regarding the content relevance and coverage of the Four Skills Exam, it is instructive to revisit the Spanish language functions which originally guided the development of the instrument. Again, Valdés (1989) states:

> Individuals surveyed agreed that teachers could communicate with young children and their parents if they were able to understand child language and both rural and standard varieties of Spanish, and if they could explain normal school requirements and activities (for example, school-yard policies, pull out programs) to persons not familiar with those concepts. Teachers were described as able to teach in Spanish if they could present material in this language easily and comfortably, without undue pauses and hesitations and without revealing large vocabulary gaps. They were able to read in Spanish if they could draw meaning from texts normally used in third grade classrooms. These texts included math, social studies, and reading materials. Finally, individuals interviewed agreed that in order to teach the Spanish writing system to young children, teachers should be able to spell correctly, use the written accent correctly, proofread material, and find mistakes in children's writing. Moreover, a teacher should be able to write notes home to parents containing few orthographical errors. (p. 218)

The intent of the test developers was to take the language skills identified by experts in the field of bilingual education in New Mexico and create language tasks that would entail relevant content and sufficient coverage in order to reliably provide an indication of the ability of the test-taker to meet the linguistic tasks.

From the above excerpt, one can infer that the aural demands of interest centered around the ability of the bilingual education teacher to understand child and adult language in both rural and standard varieties of Spanish. These aural demands are measured through the Listening Comprehension, Dictation, Informal Words, and Formal Equivalents subtests. However, only the Listening Comprehension subtest utilizes child and adult language that transcends the level of a sentence. Moreover, the Dictation, Informal Words and Formal Equivalents subtests are more like vocabulary measures than listening comprehension measures since the examinee is not required to extract meaning from the sentences but rather generate or identify a word. Furthermore, the examinees must correctly spell their answers on the Dictation and Formal Equivalents subtests in order to get the item correct. In fact, the Four Skills Exam does not appear to adequately measure the content it purports to measure in terms of listening comprehension skills.

With respect to the oral demands targeted by the Four Skills Exam, there is one glaring oversight. None of the oral tasks require the test-taker to demonstrate his or her ability to present material in Spanish. The oral tasks are geared more directly towards measuring the ability of the examinee to explain, without interruptions or exchanges, school activities, policies, etc. to adults and children. Unfortunately, one of the most salient oral skills is left unmeasured by the Four Skills Exam, the ability of the bilingual education teacher to deliver instruction in the Spanish language.

Regarding the reading skills the Four Skills Exam was intended to measure, in this case the ability to extract meaning from classroom texts, only two of the reading subtests provide such information, the Identifying Concepts and Words in Context subtests. The remaining two subtests, which comprise 40 of the sixty test items in this part of the test, are orthographic in nature. The Orthography: Accents subtest is given over completely to accentuation while the Orthography: Spelling subtest is given over to measuring the ability of the examinee to identify misspelled words. In effect, the ability of the test-taker to extract meaning from classroom texts is measured by two subtests or a total of twenty multiple choice items.

Essentially, four subtests (i.e., Dictation, Formal Equivalents, Orthography: Accents and Orthography: Spelling) are designed to measure the prospective bilingual education teacher's ability to spell correctly, use the accent correctly and proof-read children's writing. However, this fact has had the undesirable effect of reducing the number of measures to gauge the aural and reading ability of the test-taker.

Lastly, the Four Skills Exam was intended to make a judgment regarding the ability of the examinee to write notes home to parents with few orthographical errors. The writing task in the Composition part of the test appears to meet this criterion. However, one issue must be raised. How many is a few? In the case of the Four Skills Exam, a few errors is synonymous with twenty errors, an issue to be dealt with more fully in the following section regarding the technical standards of the instrument.

In summary, and in terms of the intended content relevance and coverage of the Four Skills Exam, it seems that only the Composition part of the test actually measures what the test development team set out to measure. It also appears that the aural and reading portions of the exam measure skills that are more related to orthography, primarily spelling. Again, content coverage of the aural and reading skills suffer. Lastly, the oral part of the exam failed to measure the ability of the examinee to demonstrate his or her ability to present material in Spanish. In short, at least three parts of the Four Skills Exam do not readily measure the content they were intended to. This, of course, makes it difficult to generate valid judgments about the abilities of the test-taker to meet the targeted linguistic demands which guided the development of the test.

Technical Standards of the Four Skills Exam

Evidence which can be used to gauge the unified validity of any psychometric instrument is also generated by examining the technical standards of the measurement instrument under investigation. Such standards generally provide evidence of concurrent criterion relatedness and entail a statistical examination of the reliability of the individual test items, sub-sections and the inter-correlations of the test as a whole. Similarly, these standards also subsume the procedures used, including benchmarks and scoring rubrics, to score the test. In addition, the technical standards of an instrument are also evidenced through the administration procedures governing the use of an instrument. Generally, this information is provided in the technical, scoring and administration manuals or documents designed to accompany the use of the instrument.

It should be stated from the outset that the Four Skills Exam has no technical or scoring manuals. However, there is scattered evidence which speaks to each of the three areas. The pieces are examined in turn.

In the Valdés (1989) article, the author makes an indirect reference to the establishment of the concurrent criterion relatedness of the Four Skills Exam. In this case it appears that the test development team sought to design a test that could distinguish among examinees with differing Spanish language backgrounds and abilities. If differences of test performance could be demonstrated, then this finding could aid in demonstrating that the test measures what it purports to. To this end, four distinct populations were identified and are described in Valdés (1989):

> Population I: Hispanic bilinguals native to New Mexico who had acquired Spanish at home and used English exclusively in the school setting. These persons had good oral skills in Spanish but had difficulty with the written language.

> Population II: Hispanic bilinguals from Latin America who had used Spanish in school and had acquired English as adults in the U.S. These persons had both good oral skills and good skills in the written language.
>
> Population III: Anglo bilinguals who had learned Spanish or used Spanish in a natural context (had spent time in a Latin American country). These persons usually had a good oral command of the language and could read and write well.
>
> Population IV: Anglo academic bilinguals who had studied Spanish as a foreign language and who normally did not interact with the Spanish speaking population. These individuals had marginal oral skills, limited writing skills, but could read well. (p. 219)

Valdés (1989) states that the intent was to design a test that could discriminate between those populations with limited Spanish language skills (i.e., populations I and IV) and those populations considered competent in the language (i.e., populations II and III). Valdés also hypothesizes on which sections of the test the different populations should perform well and poorly.

What is puzzling is that the process never appears to have been completed and the results reported, although the scores for the four different populations were used to establish performance standards, an issue addressed later in this section. In short, there does not appear to be any concurrent criterion related evidence for this instrument.

Regarding the reliability of the Four Skills Exam, Valdés (1989) makes a few references to the effect that the test-development subcommittee's statistician conducted item-analyses in order to modify or exclude items. Again, however, there is no technical manual documenting the reliability of the instrument. On the other hand, Cárdenas (1981), the statistician, did document the reliability of the Four Skills Exam in an unpublished manuscript. Nonetheless, there are at least four problems with the information yielded through these analyses.

First, there is no indication as to what form of the test (Form A, B, or C) the reliability analyses stem from. More importantly, no reference is made to the three different forms of the test. In effect, reliabilities were tabulated for only one form of the test and the test form is unknown.

Second, the number of test items Cárdenas tabulates reliabilities for does not match the number of items presently comprising the Four Skills Exam. For example, the reliability data reported by Cárdenas for Section I of the first part of the exam consist of 23 items; he does not specify the name of the subtest. Assuming that this section is the tape-mediated listening comprehension section of the Four Skills Exam, in its present form, this section consists of only 20 items. Additional analyses conducted by Cárdenas yield reliabilities for sections of the test consisting of 12, 15 and 25 test items. None of the sub-sections of the Four Skills Exam, in its present form, consist of these number of items.

Third, some of the reliability coefficients of the test items and sections reported by Cárdenas are relatively low. Reliability values can range from 0 to 1.0. The higher the coefficient, the more reliable the item or subtest is believed to be. For a high-stakes test of this nature, at a minimum, the items and subtests should demonstrate a reliability index of at least .90 (Davies, 1990). For example, only four items in section 1 of the first part of the exam demonstrated reliabilities above .90. The overall reliability index for this section of the exam was .649. Cárdenas (1981), however, is aware of this fact as noted in the following statement:

> It was detected that the definition of reliability used in the analysis has the undesirable property of yielding a low value of reliability whenever there is a low or a high degree of concordance among the responses of the examinees. For example, Table I shows that

> the Las Cruces group had the lowest reliability (.4570) while completing the highest percentage (89.4%) of the sections successfully. (p. 1)

Cárdenas also states that some of the test items should be reexamined, replaced or omitted. Because there is no technical manual available for the Four Skills Exam, it is not known, at least to this author, whether or not such a reexamination took place.

There are other serious discrepancies in the process followed to establish the construct validity of the instrument. No analyses appear to have been conducted which examine the inter-correlations of the subtests of the instrument. That is, there is no evidence that the different portions of the test are in high or low correlation to each other. One would expect, for example, that the aural and oral portions of the test be moderately correlated with one another since these two language skills are more closely associated with one another in terms of ability. That is, generally, an individual who speaks a language well also understands the spoken language well. Similarly, one would not expect to find a high correlation between the aural and reading sections of the exam since it is possible to read a language one cannot understand in its spoken form.

It is also critically important to note that the procedures used to establish inter-rater reliability of the subjectively scored portions of the test (i.e., the oral and composition subtests) are not mentioned in the Cárdenas document or to this author's knowledge in any other document related to the establishment of technical standards for the instrument. Benjamin and Navarrete (1992), in an unpublished document, note the absence of scoring guidelines for the subjective portions of the test. The authors state:

> While scoring is done by two proficient Spanish speaking judges, the FSE faces several drawbacks in maintaining the validity and reliability of its procedure.

> 1) Scoring Criteria: Lacking are detailed and specific criteria (scoring guides) for assessing the qualitative sections of the Oral and Composition subtests.
>
> 2) Scoring Guides: Not located are the benchmarks (sample responses) for the qualitative portions of the test that are needed to ensure reliability while scoring.
>
> 3) Scoring Ratings: According to Dr. Valdés (1989), three judges are recommended for rating the FSE. Yet, only two recognized judges are currently scoring the FSE in the state.
>
> 4) Identifying Examinees: The anonymity of the examinees is also lacking, thus increasing the likelihood of bias in scoring. Specifically, students are required to submit their name and other background information on the test cassette and test sheets thus allowing the judges to identify the name of the student, to recognize if an examinee is repeating the test and to know which university the student represents. (p. 2-3)

Clearly, this situation jeopardizes the validity of the instrument since these two sections of the test comprise half of the exam. One does not really know what constitutes a poor, weak, fair, good or very good oral protocol; similarly, one does not have an accurate or explicit depiction of what constitutes a poor, weak, fair, good or very good rating used to score the communication, appropriateness, and expression criteria inherent to the second step of grading of the composition. In other words, there are no "benchmarks" (i.e., samples of typical responses) which characterize the five possible ratings an examinee can receive on the oral or composition portions of the test. The lack of benchmarks has the potential of detracting from the reliability of these parts of the test.

An additional point must be mentioned regarding the scoring of the Composition portion of the exam. In order to move on to the final phase of the grading of the composition, the examinee must receive an average rating of at least three (Fair) on the communication, appropriateness and expression criteria mentioned above. In the final step of this composition scoring process, the scorer underlines each error in the composition. If more than 20 errors are found, the grading stops. If fewer than 21 errors are found, each error is assigned a severity value ranging from 1 to 3, with most errors generating a value of 2. There are 14 error types which are essentially grammatical in nature (e.g., punctuation, pronouns, agreement, morphology, etc.). Once values have been assigned to the errors, all the values are summed and an error score is generated. The error score must be less than 31 in order to pass the composition portion of the test. The following is a breakdown of the error score ranges: 16 or less (Very Good), 17 to 22 (Good), 23 to 30 (Fair), 31 to 38 (Weak) and 39 or more (Poor).

Recall Valdés (1989) states that a teacher should be able to write notes or letters home to parents with "few orthographical errors" (p. 218). Twenty errors is not synonymous with a few orthographical errors. The issue is that it is not known why twenty errors was selected as the cut-off point for grading the compositions.

Based on this review, the technical standards which influence the reliability and validity of the Four Skills Exam are weak and do not contribute as much as might be expected to the unified validity of the instrument. The analyses which were originally conducted by Cárdenas (1981) on the objective portions of the test do not appear to match the present forms of the test in terms of the number of test items; similarly, no clear reference is made to the form of the test analyzed and reference is made to only one form of the test. Moreover, the range of the reliability coefficients for the items and subtests are less than desirable from a psychometric point of view. Again, there is no evidence that correlations among the various test parts were ever conducted.

As concerns the subjectively scored portions of the test, there is no evidence that standardized scoring procedures have ever been developed for the oral and composition portions of the test. This is especially critical since these two test parts comprise half of the test.

In a separate and similarly unpublished document authored by Young et al., (1986), the authors set out to reexamine the reliability and validity of the Four Skills Exam. This study was based on a random sample of 100 examinees taking the Four Skills Exam for the first time. Reliability values were generated for the objectively and subjectively scored test items and subsections of the test. In the summary of this document Young et al., (1986) state:

> An analysis of a representative sample of first-time examinees showed that all parts of the test were highly reliable, with alpha coefficients ranging from .75 to .96 and a mean of .88. Correlations between the four subtests were moderate, indicating that the subtests were judging a common aptitude without duplication. (p. 15)

As with the original Cárdenas (1981) analyses, Young et al., (1986) make no reference to the form of the test from which the reliability data were generated. Moreover, no reference is made to the three equivalent forms of the test. Supposing that the analyses were conducted on test data from Form A of the exam, the following reliability coefficients were generated using the reliability formula commonly referred to as the Kuder-Richardson 20. Young et al., (1986: 13) summarize the data in Table 7 of their document.

The alpha coefficients for the four different parts of the test in this study range between .91 and .96, and as Young et al (1986) indicate, are highly reliable. The number of items used to generate the alpha coefficients, unlike the original Cárdenas study (1981) match the number of items currently comprising the Four Skills Exam. However, it is important to note that the Language subtest in Part 1 and the Comprehension subtest in Part 3 are each treated as one subtest in their respective sections. The Language subtest actually consists of two ten item subtests, Informal Words and Formal Equivalents. Similarly, the Comprehension subtest consists of two ten item subtests, Identifying Concepts and Words in Context.

Reliability* of Subtests and Components of FSE

	Alpha	Number of Examinees	Number of Items
Part 1 - Aural	.92	94	60
Listening Comprehension	.75	94	20
Dictation	.88	94	20
Language	.87	94	20
Part 2 - Oral			
Fluency, vocabulary, appropriateness	.96	79	9
Part 3 - Reading	.93	100	60
Orthography, Written Accent	.85	100	20
Orthography, Spelling Correction	.87	100	20
Comprehension	.83	100	20
Part 4 - Composition	.91	84	4
Communication, Appropriateness, Expression, Error Scores			

*KR-20

Young et al (1986) also conducted analyses yielding correlation coefficients between the four different parts of the test. The data are summarized in Table 6 of the Young et al., (1986: 13) document. As noted earlier, the correlation among subtests serves as evidence that the distinct parts of a test are not measuring the same ability or construct. The matrix indicates that the aural and oral portions of the test are more highly correlated with one another than with the reading and composition portions of the test. However, the oral and composition portions of the test are almost as highly correlated (.67) as the aural and oral portions (.71). Similarly, the reading and composition portions are not as highly correlated (.61) as the aural and oral portions (.71) of the test.

Correlation Among Subtests of FSE

	Aural	Oral	Reading	Composition
Aural	1.00 (*n* = 94)			
Oral	.71 (*n* = 74)	1.00 (*n* = 79)		
Reading	.59 (*n* = 94)	.57 (*n* = 79)	1.00 (*n* = 100)	
Composition	.60 (*n* = 80)	.67 (*n* = 69)	.61 (*n* = 84)	1.00 (*n* = 84)

Again, Young et al., (1986) describe these relationships as `moderate' and as evidence that the different portions of the test are measuring a common aptitude without duplication. Nonetheless, it is interesting to note the slightly higher correlation between the Oral and Composition subtests than was found between the Reading and Composition subtests.

The Young et al., (1986) study of the Four Skills Exam serves simply as yet another piece of information which might prove useful in gauging the unified validity of the instrument in question. However, and like the Cárdenas (1981) study, no reference is made to the form of the test analyzed and the other two alternate forms of the test are never referred to. Perhaps the most important issue related to Young et al., (1986) is the fact that their study was not explicitly intended to be a planned effort to reexamine the validity of the Four Skills Exam and to subsequently modify its contents.

The final issues to be raised concerning the technical standards of the Four Skills Exam are related to its administration. Since its development, the exam has been centrally administered by the University of New Mexico Testing Division (Pascual, 1981; Young et al., 1986; Valdés, 1989). Nonetheless, the administration of the Four Skills Exam is limited primarily to informing examinees about the exam test dates and results as well as the duplication, mailing-out and storage of the exam.

It is important to note that these functions, while seemingly mundane, serve important purposes which affect the technical standards of the exam. For example, the responsibility of mailing out the needed copies of the test to different test sites also entails the need to be cognizant of the form of the test (i.e., Form A,B, or C) which must be delivered in order to maintain the reliability of the test. However, if the staff member fulfilling this task has not been advised of this process, the rotation of the alternate forms of the test can easily be overlooked. In addition, it is important to state that these functions have been carried out for over thirteen years in the absence of any meaningful financial resources (Griñe, personal communication, 1993).

On a similarly dismal note, the administration manual which is critical for maintaining the standardization and reliability of the instrument, consists of general statements. This is especially critical given the fact that the test is offered three times each year at different sites across the state of New Mexico. As Navarrete and Benjamin (1993) state:

> ...a review of the administration manual reveals that many of the instructions to the test administrator are not adequately detailed. For example, there are no instructions for the administrator to introduce the test, to describe the subtests, and to explain the time period required to complete each part of the test. In other instances, subtests lack explicit instructions for the administrator to orient examinees to the test as is common in other "standardized" tests. (p. 2)

Benjamin and Navarrete (1993) also raise the issue of training the test administrators. Their review of the issue indicates that there is no explicit training plan in place.

These same authors also bring attention to an issue regarding the administration of the aural section of the exam. Valdés (1989) indicates that this portion of the exam is designed to be administered in a language laboratory; Navarrete and Benjamin indicate that such facilities are unavailable at some test sites. Valdés also indicates that the oral portion of the test is to be administered in a language laboratory; Hannum (1993b) informs examinees to bring their own tape recorder for this part of the test. Recall that these two portions of the test comprise half of the test.

With respect to the technical standards of the Four Skills Exam, all indications are that the procedures used to establish and maintain the reliability of the test do not enhance the unified validity of the test. Similarly, the scoring criteria for the subjective portions of the exam appear to exist primarily in the minds of the two official scorers of the test; moreover, there is no evidence of established inter-rater reliability. In addition, the administration manual of the test outlines general instructions which do not appear to be followed across all test sites. This is especially true in the case of the aural and oral portions of the test. Lastly, there is no evidence which indicates that the test administrators receive any type of training. Again, some of the above deficiencies are likely to have their roots in the lack of financial support.

Spanish Language Proficiency Testing in the Southwest

Each state in the Southwest (i.e., New Mexico, Arizona, California, Colorado and Texas) has designed or adopted a Spanish language proficiency test which prospective bilingual teachers seeking a bilingual endorsement must pass within their respective states. Of course, this practice is not exclusive to the Southwest as McFerren et al., (1988) indicate that a total of 18 states require the prospective bilingual education teachers to demonstrate their non-English language proficiency via a formal language assessment.

As concerns the unified validity of the Four Skills Exam, no studies have ever been undertaken to demonstrate concurrent validity among any of the Southwestern measures. That is, while each test is designed to measure some of the same constructs (e.g., oral language proficiency) for a similar purpose (i.e., a bilingual endorsement), it is not known whether an examinee would perform in a like manner on two or more of these tests. This is a critical missing piece of empirical evidence that could enhance the construct validity of the tests being compared.

Nonetheless, it is useful to at least compare the Four Skills Exam with the exams used in Texas, California and Arizona. Note that Colorado has adopted the Four Skills Exam for endorsement purposes. This comparison can yield information which may be useful at least in gauging the content relevance and coverage of the Four Skills Exam.

In Texas two tests are used for certification purposes, the Texas Oral Proficiency Test (Stansfield et al., 1991) and the ExCET Bilingual Education exam (Texas Education Agency, 1988). This latter exam actually measures the examinee's knowledge of bilingual education theory, pedagogy and Spanish language arts. Only the Spanish language arts section of the ExCET is examined in this comparison. Arizona uses the Spanish Language Proficiency Test for Bilingual Teachers developed by Riegelhaupt et al., (1981). In California the Bilingual Certificate of Competence Spanish language subtest developed by Bilingual Testing Services (1984) is used for certification purposes of bilingual education teachers already possessing an elementary or secondary teaching license.

Each test purports to be education, job, or classroom related, valid for bilingual teachers in grades K through 12, and takes about the same amount of time to administer, between two and two and a half hours. However, the Texas Oral Proficiency Test (Stansfield et al., 1991) is only used to certify bilingual teachers in grades K-6. As noted above, prospective bilingual teachers in Texas must also take the ExCET exam for bilingual education which contains a subtest for measuring Spanish literacy skills. Because Texas uses two different tests, it is difficult to ascertain the amount of time needed to administer the tests.

Given this initial information, the Arizona exam appears to be most like the Four Skills Exam. It is also instructive to note that both of these exams were first used in 1981 and both purport to be based on `real life' bilingual education teacher language skills (Riegelhaupt, 1985; Valdés, 1989). In fact, both authors indicate that the test development process entailed careful observations of the uses to which Spanish was put in bilingual classrooms in each respective state. In this sense, neither the Texas or California purport to be based on `real life' bilingual education language skills.

Each test requires the examinee to demonstrate oral proficiency through a tape-mediated format. However, the Four Skills Exam appears to demand the least from the examinee. First, the examinee's score on the oral portion of the Four Skills Exam can be based on as little as three minutes of discourse since the examinee only needs to earn a passing score on two of the three tasks. In contrast, the Texas instrument generates a twenty minute sample of oral discourse, while the Arizona exam generates approximately a ten minute sample and the California exam an eight minute sample of oral discourse. The length of time is important since longer tests are more likely to be more reliable than shorter ones (Bachman, 1990).

In addition, the focus of the oral section of the Four Skills Exam is clearly on the examinee's ability to explain a situation to a child or adult as noted earlier. The other three exams go beyond the explanatory function. For example, the Arizona exam consists of four sections each of which require the examinee to demonstrate different language functions: to read a passage aloud, to conduct an instructional activity, to formulate questions based on a reading and to communicate with parents. It is also evident that the Arizona exam integrates language abilities through its test format. For example, the examinee must `read' a passage aloud on one subtest, and orally formulate questions based on a reading on another subtest.

The Texas instrument consists of three scored sections each of which requires the examinee to respond to five pictorial or written prompts. In other words, the examinee responds orally to fifteen different prompts which vary in their degree of complexity. For example, the examinee may be asked to order a meal, describe a sequence of events in the past or give a professional talk. The California exam requires the examinee to engage in three distinct oral activities: a response to a hypothetical situation, responses to five questions or tasks asked in Spanish, and a reading aloud task.

With respect to the aural domain of language proficiency, neither the Texas or California exams generate a measure of this ability. Recall that this section of the Four Skills Exam consists of four tape mediated subtests: Listening Comprehension, Dictation, Informal Words, and Formal Equivalents.

The Arizona exam contains two sections which generate a measure of aural comprehension. In section (1) of the exam, the examinees watch and listen to a video of two children verbally interacting in a classroom. The examinee must answer multiple choice questions based on their comprehension of the children's verbal exchanges. In section (6) of the same exam, the test taker must listen to a contrived telephone conversation with a parent and respond appropriately to the parent's verbal prompts. Again, the `real life' nature of this test is evident. The visual sensorial input which is common to human communication (Oller, 1991), while not direct, is at least present in one section of this test format. Similarly, engaging in a conversation, though contrived, requires both listening and speaking skills, a natural combination of language skills.

With respect to its oral section, the Four Skills Exam offers limited content relevance and coverage when compared to these other tests. It is interesting to note that only the Arizona exam generates a measure of the examinee's ability to deliver instruction in Spanish, a central function in bilingual education. In this way the Texas and California exams also fall short. In addition, neither of them, makes any effort to measure listening (i.e., aural) ability. Clearly, bilingual education teachers need to be able to understand spoken Spanish discourse.

However, mere inclusion of aural comprehension subtests does not automatically assure behavioral relevance. For reasons stated earlier, the Four Skills Exam aural section does not appear to be authentic. Arizona's aural subtests, through the use of video and a contrived telephone conversation, appear to be most behaviorally relevant, to possess a greater degree of content relevance.

In the domain of reading abilities, recall that the Four Skills Exam consists of four subtests: Orthography: Accents, Orthography: Spelling, Identifying Concepts and Words in Context. The Texas ExCET exam and the California exam both employ the traditional reading comprehension format much like the third and fourth sub-sections of the Four Skills Exam. None of the other three tests, however, utilize actual text book excerpts as readings in these sections of the exams. On the other hand, none of the other three tests measure accentuation or spelling correction as reading abilities.

It is worth noting that only the Arizona exam again utilizes what could arguably be construed as more `real life' reading tasks which integrate language skills. For example, one of the subtests require the examinee to read an authentic text from a professional journal written in Spanish. The examinee must then write a brief summary of the passage in Spanish. Similarly, and already mentioned above, on another subtest the examinee is required to read a short passage and to then formulate questions based on the text. The examinee then orally records the questions.

Concerning the prospective teacher's writing ability, recall that the Four Skills Exam requires the examinee to write a letter, at least 150 words long, to parents on a designated topic. The ExCET exam used in Texas assesses writing skills through multiple choice type test items; no separate writing score is generated, however. California uses a format similar to New Mexico's but in this case the examinee is required to write two one page essays on designated topics. Again, the issue of reliability can be raised since the score of the test-taker on the California exam is based on a larger writing sample than is the case for the New Mexico examinees.

Arizona, as noted above, integrates reading and writing skills. Recall that the examinee must read a professional text and then summarize the passage in writing. Again, one could question the 'real life' nature of such a task. Bilingual educators would seem more likely to read such texts but to then discuss them orally with their peers not summarize them in writing. On the other hand, the Arizona exam also requires the test taker to translate an official school document from English to Spanish, a potentially relevant task. Lastly, this exam requires the examinee to proof-read a student's composition and to correct potential writing errors, an ability measured by the Four Skills Exam in its reading section through the decontextualized spelling correction subtest.

In fact, the Four Skills Exam appears to be the only test in the Southwest which includes accentuation and spelling correction under reading ability. On the other hand, New Mexico's exam is the only exam which uses authentic bilingual program reading texts in this section of the exam. Nonetheless, the Four Skills Exam does not appear to integrate language skills in a real life manner as is the case in Arizona's exam. The writing demands placed on the examinee on the Four Skills Exam, writing a letter to parents, appears to be as 'real life' as the translation of an official school document as required in the Arizona exam.

This brief comparison of the content relevance and coverage of the Four Skills Exam with the three exams currently in use in the Southwest points to potential weaknesses and strengths of the instrument under consideration. The greatest potential weakness is clearly in the oral section of the test due to its brevity and narrow range of language functions. The aural section, while mainly discrete point and based on disconnected discourse unrelated to a bilingual education setting, at least acknowledges the importance of this ability. The Texas and California tests do not measure this ability. In this sense, the aural content of the Four Skills Exam adds to the validity of the instrument. On the other hand, the aural component of the Arizona exam appears to be more `real life' than the aural component of the Four Skills Exam.

One of the strengths of the Four Skills Exam lies in the use of authentic reading excerpts from actual text books used in bilingual programs. None of the other three tests acknowledge their use in their respective instruments. On the other hand, accentuation and spelling correction appear misplaced as part of the reading portion of the Four Skills Exam. Moreover, none of the other tests place such a heavy emphasis on accentuation.

In terms of writing ability and content relevance, it seems that the composition portion of the Four Skills Exam is more valid than either of the written portions of the California or Texas exams since the New Mexico exam is based on a potentially real life writing demand. On the other hand, the writing score of the California exam is based on a larger writing sample than the single composition required by New Mexico. Lastly, the written portion of the Arizona exam does not necessarily appear any more valid than the corresponding section of the Four Skills Exam. Both exams have identified at least one important `real life' writing skill, but neither include both writing a letter home to parents and the translation of a school document.

This brief comparison demonstrates that it is feasible to design a language proficiency test that utilizes visual and kinesic semiotic stimuli such as pictures, diagrams and video. Similarly, this review demonstrates that it is also feasible to integrate language skills in a behaviorally relevant and contextualized manner as accomplished in the Arizona exam. Both of these desirable test qualities are lacking in the Four Skills Exam. In terms of content relevance, it appears that each test is lacking in one manner or another. However, only the Four Skills Exam places such a heavy emphasis on accentuation and spelling. With respect to content coverage, the aural, oral, and reading measures used to assess these skills in New Mexico seem limited in terms of the number of test tasks given over to the measurement of these abilities. On the other hand, the Four Skills Exam does have some qualities lacking in the California and Texas exams.

Test Validity In Social Context

There is an interesting irony to the Spanish language proficiency testing policies in New Mexico, and throughout the U.S. in general, which merits addressing. On the one hand, state adopted competencies like those in New Mexico call for fairly high levels of Spanish language proficiency, including literacy. Moreover, a language testing policy is in place to attempt to ensure that prospective bilingual education teachers can at least meet the linguistic criteria inherent to an adopted language proficiency test. While it appears to make sense to have such a language testing policy in place, there is an undeniable paradox.

Many of the social institutions (e.g., the educational, political, judicial, economic systems, etc.) in the U.S., including New Mexico, discourage the maintenance of the Spanish language and bilingualism in general, especially among native speakers of the Spanish language. Numerous authors describe how the educational experience in the U.S. is a monolingual English one and how non-English proficiency is viewed as a deficit (August & Garcia, 1988; Crawford, 1989; Cummins, 1989; Kjolseth, 1983; Lyons, 1990; Ruíz, 1988). Similarly, other authors have developed arguments which demonstrate how the political, judicial and economic systems in the U.S. perpetuate English monolingualism and disparage the maintenance of non-English languages (Hernández-Chávez, 1988; Gonzáles et al., 1988; Grenier, 1984; Peñalosa, 1980; Piatt, 1990; Spener, 1988; Tienda & Neidert, 1984).

The impact this xenophobic social infrastructure has had on the maintenance of Spanish in this country is also well documented. There are numerous studies that document the process of language shift and loss which characterizes the linguistic experience of the vast majority of the potentially bilingual Spanish-English speakers in the U.S. (Bills, 1989; Solé, 1990; and Veltman, 1988).

Ada (1986) succinctly summarizes the negative effect the linguistic milieu in the U.S. can have on bilingual education teachers in the following paragraph:

> Bilingual teachers may feel inadequate in their language ability because of several factors. Those teachers whose mother tongue is English may not have had the opportunity to acquire full mastery of a second language- a sad reflection on our limited and deficient foreign language teaching. Members of language minorities who chose to become bilingual teachers may also have been victims of language oppression as children, when they were scolded or punished in school for using their home language. Therefore it should not be surprising that many bilingual teachers lack confidence in their literacy skills. Yet if these individuals can acknowledge that the language inadequacy they experience stems from deeply rooted institutionalized oppression and

> is high-lighted by the one-teacher model, they will be better able to understand what their students may be going through. (p. 390)

The sociolinguistic context in the U.S. in general is undeniably one which only embraces English monolingualism. Nonetheless, prospective bilingual education teachers, especially members of language minority communities, are expected to escape the wrath of deeply rooted institutionalized oppression or linguicist policies (Phillipson, 1988; Skutnabb-Kangas, 1990) and develop language skills they have historically never had access to (Hernández-Chávez, 1993). In keeping with Messick's unified validity framework, primarily the consequential basis of test use, one must ask how this sociolinguistic variable impacts the validity of the Four Skills Exam .

Summary: Review of the Literature

Given this review of the literature, ample evidence which is needed in order to move towards setting forth an overall evaluative judgment of the validity of the Four Skills has been generated. The evidence is presented so as to address the four major areas of validity (i.e., construct validity, content relevance and coverage, value implications of score interpretation and the social consequences of score interpretation) advocated by Messick (1989).

I. Construct validity

The balance of evidence which would support the construct validity (i.e., behavioral relevance) of the Four Skills Exam is limited. Valdés (1989), which is essentially the sole source of information on the exam, makes no explicit reference as to the theoretical model of language which guided the development of the test. Valdés does, however, indicate that the test development team sought to develop an instrument based on "real life" bilingual education teacher language demands. Nonetheless, the behavioral relevance of the Four Skills Exam, when compared to the semiotic model of language proficiency advocated by Oller (1991), does not readily reflect real life language processing. The lack of behavioral relevance is also evident based on the comparison of the Four Skills Exam to the Arizona exam. Apparently, the test development team failed to fully operationalize the constructs to be measured (e.g., speaking, listening, reading and writing).

Additional evidence which bears on the construct validity of the Four Skills Exam (i.e., criterion relatedness, reliability analyses, and correlation analyses) is piece-meal and basically unavailable. The Cárdenas (1981) and Young et al., (1986) studies examine an unknown form of the test. Similarly, empirical research aimed at establishing the concurrent validity and consequently the construct validity of the Four Skills Exam has not been undertaken.

II. Content relevance & coverage

The content relevance and coverage of the Four Skills Exam, when evaluated based on the original intent of the test development team, also does not lend support to the overall validity of the Four Skills Exam. While the test development appeared to have identified the Spanish language functions which would form the essence of the test, they failed to adequately embed the target language functions into the test, especially in the case of the aural, oral and reading portions of the test. Either the content of the subtests comprising these sections of the test did not match the targeted content or the content was not used at all. This oversight also led to marginal content coverage. On the other hand, the test developers do appear to have adequately addressed content related to spelling, accentuation, and composition given their original intent.

The content relevance and coverage of the Four Skills Exam when viewed from the perspective of the other three tests used in the Southwest is especially weak in the assessment of oral language proficiency. This part of the test is limited in the oral language tasks assessed, especially as concerns the ability of the examinee to deliver instruction in Spanish. Furthermore, the length of time on which the oral ability of the examinee is based is brief in comparison to the other measures reviewed.

Valdés (1989), while not explicitly stating so, gives the indication that the Four Skills Exam was intended to measure the kinds of language skills bilingual education teachers in New Mexico would need in order to teach young children, perhaps between kindergarten and fourth grade. If the test development team was only partially successful at embedding the content relevant to the early grades, clearly the content relevance and coverage for grades five through twelve is even more suspect. Again, this is not the fault of the test developers, but rather the policy makers responsible for using the Four Skills Exam for a purpose for which it was never intended.

Based on the New Mexico State Department of Education native language competencies for bilingual education teachers, the content relevance and coverage of the Four Skills Exam is inappropriate for making a determination as to whether or not a bilingual education teacher has met these competencies. Again, the test was not developed to measure the content entailed in these competencies.

III. Value implications of score interpretation: inferences derived from test scores

Given the evidence related to the construct validity, reliability, and content relevance and coverage of the Four Skills Exam, it would appear difficult to make valid judgments regarding the Spanish language classroom-related abilities of a prospective bilingual education teacher in New Mexico. The aural subtests, which one would assume are intended to determine whether or not the teacher candidate can understand the standard or rural Spanish spoken by children and adults, do not readily measure this ability. Only the Listening Comprehension subtest taps this ability. Consequently, making a valid inference about the aural comprehension abilities of the test-taker in a Pass or Fail manner is difficult. Some individuals who pass this portion of the test may not be able to comprehend spoken Spanish; others who fail it may understand spoken Spanish natively. The inference that can drawn from a pass or fail score on the reading portion of the exam is likely to be as dubious since two of the reading subtests are given over to aspects of orthography. The judgment regarding the ability of the examinee to extract meaning from a school related text is based on two ten item multiple-choice subtests. Again, some individuals who pass this portion of the test may have good orthographic skills, but not necessarily adequate reading comprehension skills and vice versa.

The meaning of a pass or fail score on the oral part of the exam is also elusive. As stated previously, there is no measure of the ability of the examinee to deliver instruction in Spanish. Moreover, there are no clearly established benchmarks for scoring the oral discourse samples. To agitate matters even more, the examinee can pass this portion of the exam by scoring "fair" on only two of the three passages. In effect, the test taker can be deemed orally proficient in classroom Spanish based on a three minute sample.

The meaning of a pass or fail score on the composition part of the exam is also open to interpretation. As with the oral part of the exam, there are no benchmarks to guide the scoring of the composition. In addition, the real meaning of a pass score on the composition could mean that the test-taker made up to twenty errors within a text consisting of 150 words. Can an individual who makes twenty, fifteen, ten or five errors effectively teach the Spanish writing system to young children? More importantly, it seems unlikely that English speaking mainstream classroom teachers would be held to such a low standard.

IV. Social consequences of using scores for applied decision making

The social consequences of using the scores of the students generated by their performance on the Four Skills Exam are straight-forward. Some of the bilingual education teachers who have passed the test may not have the language skills they or others believe they do. Obviously, these individuals may be unable to meet the educational needs of the Spanish speaking students they are responsible for. On the other hand, other bilingual education teachers who have failed the test may in fact have the Spanish language abilities the test development team originally intended to measure. In either case, the students who stand to benefit from bilingual instruction suffer the greatest consequences.

The responsibility for this dismal Spanish language testing enterprise for bilingual education teachers in New Mexico lies with several parties. The original test development team seems to have met with some language test development challenges they were not fully able to overcome. However, the establishment and documentation of the reliability and scoring procedures for the instrument seems to be a gross oversight for such a high-stakes exam. The New Mexico State Department of Education has also contributed to the situation by endorsing the use of the Four Skills Exam for purposes for which it was never designed. In addition, while the state department readily endorses the (mis)use of the Four Skills Exam, this office has not offered any financial support which is essential for maintaining the validity of any testing process.

Additional evidence is forthcoming in the remaining chapters. Chapter Three, which follows, sets the stage for the statistical analyses which were needed in order to move closer towards the goal of this dissertation, an overall evaluative judgment of the unified validity of the Four Skills Exam.

CHAPTER 3

METHOD

In this chapter a profile of the subjects whose test scores were used to carry out this research is described. The procedures used to create a common metric among the four distinct parts of the test and to quantify sociodemographic data are also described. The statistical analyses used are also identified in relation to the central questions researched in this dissertation.

Subjects Sociodemographic and other relevant data for each examinee were available from the Cover Sheet (Hannum, 1993a) of the test booklet and the Four Skills Exam Official Score Report. The subjects involved in this study were 217 examinees ($n = 217$) who took the Four Skills Exam for the first time between 1991 ($n = 75$) and 1992 ($n = 142$). Sixty seven percent ($n = 146$) of the examinees took Form A, 28.1% ($n = 61$) took Form B, and only 4.6% ($n = 10$) of the examinees presented Form C of the exam. The distribution regarding the test site where the examinees took the test is summarized in Table 1.

The category Other in Table 1 refers to those examinees that took the exam at a test site which was not affiliated with an institution of higher education.

Table 1

Test Sites Across New Mexico

Test Site	Number of Examinees	Percent of Examinees
Site 1	68	31.3
Site 2	50	23.0
Site 3	31	14.3
Site 4	37	17.1
Site 5	14	06.5
Other	14	06.5
Missing	03	01.4
Total	217	100

With regard to the ethnicity of the examinees and based on the judgment of this author, 74.2 % (*n* = 161) of the examinees had a Hispanic surname and 25.8 % (*n* = 56) had a non-Hispanic surname. With regard to where the examinees had grown up, the majority (69.1%) of the examinees were native to New Mexico. Table 2 summarizes the relevant data:

Table 2

Geographic Location Where Examinees Grew Up

Geographic Location	n	Percent
New Mexico	150	69.1
Southwest	20	09.2
Spanish Speaking Country	11	05.1
Elsewhere	29	13.4
Missing	07	03.2
Total	217	100

With respect to where the examinees reported residing, there was a fairly even geographical distribution of examinees across the northern (i.e., from the New Mexico and Colorado border to Santa Fe), central (i.e., from Bernalillo to Socorro) and southern region (i.e., all towns and cities south of Socorro) of the state of New Mexico. Exactly 34.6 % (*n* = 75) of the examinees reported residing in the northern region, 34.1 % (*n* = 74) in the central region and 25.3 % (*n* = 55) in the southern region. 4.6 % (*n* = 10) reported living outside of New Mexico.

In terms of the Spanish language background of the examinees, 22.1 % (*n* = 48) reported not speaking Spanish either as they grew up or presently at home. 24.4 % (*n* = 53) reported speaking Spanish either as they grew up or presently at home; while the majority of the examinees, 53.5 % (*n* = 116) indicated that they grew up speaking Spanish and presently speak Spanish at home.

Only 6.5 % (*n* = 14) of the examinees reported not having studied Spanish formally in high school or college, while 37.3 % (*n* = 81) indicated that they studied Spanish either in high school or college; the majority, 56.2 % (*n* = 122), reported having studied Spanish in high school and college.

These data also indicated that 59.4 % (*n* = 129) of the subjects were teachers; 3.7 % reported having a teacher aide status and 31.8 % (*n* = 69) a university student status. 4.6% (*n* = 10) reported having some other type of status. 37.8 % (*n* = 82) of the examinees were not teaching in a bilingual program, while 54.8 % (*n* = 119) reported teaching in a bilingual program at the time of taking the exam. There were 16 missing cases for these data. The distribution of the examinees across the different grade levels in which they reported teaching are summarized in Table 3.

The composite of the average examinee given these data is as follows. The average examinee is likely to have a Hispanic surname and to have grown up in New Mexico in any of the three geographic regions of the state. The examinee is also likely to have grown up speaking Spanish and continues to speak Spanish at home. In addition, this individual has had some formal college Spanish language course work and presently teaches in a bilingual program at the elementary school level.

Table 3

Grade Level Distribution of Examinees

Grade Level Taught	Number of Examinees	Percent
Elementary	98	45.2
Middle School	20	09.2
High School	17	07.8
Institute of Higher Education	02	00.9
Other	05	02.3
Missing	75	34.6
Total	217	100

Materials The tests used in this study were taken at one of five test sites around the state. All tests had been completely hand scored by the two officially designated test scorers in the state. Further, the tests used in this study were housed at the Testing Division at the University of New Mexico. Test data were input onto the mainframe 9121 computer at the Computer and Information Resources and Technology Center also located on the university's campus. Statistical analyses were conducted using the Statistical Package for Social Sciences or SPSS-X (1988).

Procedure For the purposes of examining the reliability of the instrument and answering the set of questions central to this dissertation, the following general procedures were followed.

Each examinee was assigned an identification number which in most cases corresponded to their social security number. Then, each examinee's test answer selections were numerically recoded and entered into the mainframe computer. Specific-ally, Part 1 (Aural) and Part 3 (Reading) of the three forms of the exam (i.e., Forms A,B, and C) had to be numerically recoded since the test format for these two parts was multiple-choice and fill in the blank. Consequently, multiple choice options A,B, C, and D were numerically coded as 1, 2, 3 and 4, respectively. Test items which were dichotomous (i.e., fill in the blank) were assigned a numerical value of 0 for incorrect answers and 1 for correct answers as determined by the official scorers.

It was also necessary to rescore the multiple choice sub-sections of these two parts of the exam across each of the three forms of the test in order to conduct the reliability analyses. The correct answers for these sections of the exams were ascertained by examining the test booklets of those examinees who successfully answered each test item as determined by the official test scorers.

As previously explained in Chapter 2, Part 1 (Aural) consisted of four tape-mediated subtests: 20 multiple-choice Listening Comprehension items, 20 fill in the blank Dictation items, 10 fill in the blank Informal Word items and 10 fill in the blank Formal Equivalent items. Part 3 (Reading) also consisted of four subtests: 20 multiple-choice Orthography: Accents items, 20 fill in the blank Orthography: Spelling items, 10 multiple-choice Identifying Concepts items and 10 multiple choice Words in Context items. Each part consisted of sixty items, and the examinees' responses were tabulated for each subtest and the parts of the test as a whole.

Test scores for the subjective portions of the exam, Part 2 (Oral) and Part 4 (Composition) were available from the FSE Official Score Report. These two parts of the exam were not rescored.

Some numeric transformations were necessary in order to give each part of the exam equal weight (i.e., sixty points). Part 2 consisted of three oral passages each of which were rated for being on topic, fluency, vocabulary and appropriateness on a rating scale consisting of Poor, Weak, Fair, Good and Very Good.

It is important to mention that the criterion 'On Topic' is a yes or no decision made by the scorer. If the examinee is On Topic, scoring continues; if the examinee does not address the stipulated topic, scoring stops. In assigning numeric values to this criterion, the value was either five for being On Topic or zero for not being On Topic.

The rating scale descriptors for fluency, vocabulary and appropriateness were replaced with numerical values ranging from one (1) to five (5), one (1) being equal to a Poor rating, and two (2) to a Weak rating, and so forth up to five (5) for a Very Good rating.

These transformations allowed for the generation of a numerical value between zero and twenty for each passage. Since each passage was worth no more than twenty points, and the section consisted of three passages, a score of sixty was also the highest possible for this part of the exam.

It is also important to mention that the numerical ratings for each of the four criteria scored (i.e., on topic, fluency, vocabulary, and appropriateness) on the Oral part of the test were input into the data set. In this way, it would also be possible to examine the internal consistency of the ratings generated by the two scorers.

The test scores from Part 4 (Composition) of the exam underwent a similar conversion. The scoring of the composition may consist of one to three steps. In Step One, the test scorer makes a determination as to whether or not the examinee wrote on the designated topic and if the composition contains a minimum of 150 words. Only if these two criteria are met will scoring of the composition move on to Step Two.

For the purposes of data analysis, if these criteria were met, a numerical value of one was entered for each criterion; if either of the criteria were not achieved, then a value of zero was input into the subject's record.

In Step Two of the scoring of the composition, the scorer rates the composition on three separate criteria: communication, appropriateness and expression. Again, the ratings range from poor, weak, fair, good and very good. These verbal descriptors were assigned a numerical value from one to five as in the Oral section of the exam described above. These three numerical values were also individually entered into the examinee's record in order to examine the internal consistency of the ratings given by the two scorers across the three criteria.

BEST COPY AVAILABLE

An average score of three is needed in Step Two of the scoring of the composition in order for the scoring to proceed to the final and third phase of scoring. In this final step, the scorer underlines each grammatical error found within the first 150 words of the composition. If more than twenty errors are found, the scoring stops. If fewer than twenty errors are found, then the scorer assigns error points, ranging from one to three points, for each error.

Finally, the error points are totaled and a final rating ranging from weak to very good is assigned to each composition, depending on the error score. Again, a numerical value was assigned to each verbal descriptor with a Poor rating equal to one and a Very Good rating equal to five, as previously described. For those examinees who did not get past Step One of the scoring or received less than an average rating of three on Step 2 of the scoring procedures, a final rating of zero was entered into the examinee's data record.

In essence, each examinee receives a final rating of zero to five on the composition. For the purposes of creating a common metric across all four parts of the test, the final rating was simply multiplied by twelve. Consequently, the score of an examinee can range from zero to sixty.

In order to conduct the analyses, numerical values were generated for each item in each subtest and in Parts 1 (Aural) and 3 (Reading). Similarly, numerical values were created for the target criteria on Parts 2 (Oral) and 4 (Composition) as well as for the parts as a whole. Lastly, each of the four parts of the exam was worth a maximum of sixty points with a total score possible of two hundred and forty.

Two additional variables were created in order to conduct the needed analyses related to the equivalency of the test forms. These two variables subsumed those parts of the exams which were either common (i.e., Listening Comprehension, Informal Words, and Formal Equivalents subtests, Oral Passage One, and Part 4 (Composition) or uncommon (i.e., the Dictation subtest, Oral Passage Two and Three, the Orthography: Accents, Orthography: Spelling, Identifying Concepts and Words in Context subtests) to each of the three forms of the test. The differences in the three forms of the test would obviously be attributed to the uncommon parts variable, if differences were found. As noted previously in Chapter 2, Oral Passage One was exactly the same across each of the three forms of the test and for this reason was considered a common element. In the case of Part 4 (Composition), the examinee is given two topics from which to choose and must write on only one of the topics. Across each of the three forms of the test, the first topic concerned writing a letter to parents requesting permission for the student to participate in a school field-trip; the second topic also concerned writing a letter to parents. Since both writing tasks focused on a letter to parents, this part of the exam was considered common across the three forms of the test.

Regarding the procedures used to statistically analyze the examinees' test performance as it related to sociodemographic variables, the following procedures were followed. Again, sociodemographic information was taken from the FSE Official Score Report and Cover Sheet.

Some of the data collected readily lent themselves to numeric representations (i.e., test form, teaching level, teaching in a bilingual program, where the subject grew up, and yes/no answers to questions regarding their Spanish language background). For example, the forms of the test (i.e., A,B, C) were given values of one, two, or three, respectively. Where the subject grew up was either in New Mexico, the Southwest, a Spanish speaking country or elsewhere. Hence, numerical values ranging from one to four were then created.

The four questions regarding the examinees' Spanish language background were condensed into two values ranging from zero to two. For example, those examinees' reporting having spoken Spanish as they grew up and presently at home received a two value; those answering only yes to one of the questions received a one rating, while those answering negatively to both questions received a zero rating. A similar procedure was used for numerically representing the test-takers' formal studies in Spanish.

There were two cases in which some subjective decisions had to be made. The first case involved the subjects' surname. This researcher, using his best judgment, categorized each subject's surname as either Hispanic surnamed or not Hispanic surnamed.

The second case in which some subjective judgment was used concerns the examinees' geographic residency. The `City' in which the subjects' reported residing was geographically categorized as either northern, central or southern. All cities south of Socorro were considered southern; cities between Socorro and Bernalillo were deemed central; cities north of Bernalillo were classified as northern.

It is also important to mention that the variable `test site' was used inferentially to determine the examinees' institutional affiliation. No data were available which directly linked the examinee to the institution where the subjects received their bilingual education teacher training. The assumption, thus, was that the test site (i.e., one of five institutes of higher education) where the examinees' took the exam was the same institution where they received their training. At any rate, all of the sociodemographic data needed for this dissertation were converted to numeric values. Some of the similar items had to be condensed and, in a few cases, inferential reasoning was necessary to make use of the somewhat limited sociodemographic data available. In fact, it should be kept in mind that the author of this dissertation did not design the demographic questions nor have any control over their application at the test sites.

Analyses In order to generate additional evidence which would allow for an overall evaluative judgment of the unified validity of the Four Skills Exam, several statistical analyses were conducted. For the purposes of examining the reliability of the exam, and ultimately the construct validity of the instrument, Cronbach's α was employed. Reliability coefficients were calculated in order to examine the internal consistency of each subtest. In addition, product-moment correlation coefficients (r) were also generated using the Pearson formula for all the parts and subtests of the Four Skills Exam. Lastly, in order to test for the equivalency of forms, specifically the equivalency of the uncommon parts variable, analyses of variance (ANOVA) were used.

One of the most fundamental questions addressed in this dissertation is: What was the Pass and Fail rate of the examinees on the different parts of the test and the test as a whole? Descriptive data were generated which provided the percentages of the examinees passing and failing each part of the exam and the exam as a whole.

A series of multivariate analyses of variance (MANOVA) were also conducted in order to generate evidence which would also provide insight into the construct validity of the test.

Statistical significance was set at the .05 level. Questions addressed using these analyses were:

1. Did the test performance of the examinees vary as a function of formal language training?
2. Did the test performance of the examinees vary as a function of Spanish language background?

ANOVA and MANOVA analyses were also conducted in order to answer the following questions related to the social consequences of using the exam:

3. Did the test performance of the examinees vary as a function of institutional affiliation?
4. Did the test performance of the examinees vary as a function of the region of the state they happened to reside in?
5. Did the Hispanic surnamed examinees perform differently from the non-Hispanic surnamed examinees?

CHAPTER 4

RELIABILITY ANALYSES AND RESULTS

In this chapter statistical results using Cronbach's α are reported which provide evidence of the reliability (i.e., internal consistency) of the Four Skills Exam. Language researchers concur (Bachman, 1990; Cohen, 1994; Davies, 1990) that internal consistency is concerned with how consistent an examinee's performances on the various parts of a test, including individual test items, are with one another. Cronbach's α generates a coefficient alpha ranging in value from 0 to 1.0. The closer the coefficient alpha is to 1.0 the more internally consistent the test item or subtest is believed to be. There is no clear consensus among language testing experts as to what constitutes an acceptable level of reliability. Hughes (1990) suggests that an item showing a correlation of (0.3) is satisfactory. Davies (1990) indicates that a test must have a correlation coefficient of at least (0.90). On the other hand, Cohen (1994) maintains that standardized tests used for large scale administration should demonstrate reliability coefficients of at least (0.80). Because the Four Skills Exam is a high-stakes test which examinees are required to pass only once in their professional career as a bilingual education teacher, the former standard, (.90), is used as an acceptable level of reliability for the purposes of this dissertation.

Coefficient alphas for individual test items and subtests common to forms A, B and C of the Four Skills Exam are reported in Tables 4 through 8. The Common Parts are: Listening Comprehension, Informal Words, Formal Equivalents, Oral Passage One, and the Composition. The reliability results of the test items and subtests for the Uncommon Parts of the test for Form A are provided in Tables 9 through 15 and for Form B in Tables 16 through 22. The Uncommon Parts are: Dictation, Oral Passages Two and Three, Orthography: Accents, Orthography: Spelling, Identifying Concepts, and Words in Context. The reliability coefficients for all the test parts are summarized in Table 23.

Due to the small number of cases (*n* = 10) in which Form C of the exam was used, reliability analyses could not be meaningfully conducted for this form of the Four Skills Exam. However, Form C data were used in the common parts analyses. Lastly, interpretations of the analyses follow each set of tables germane to each of the four parts of the exam.

Common Parts

Table 4 below indicates that the Listening Comprehension subtest has a relatively low level of reliability in terms of its internal consistency (.5467). Note that items 4, 8, 9, 13, 14, and 15 contribute minimally or negatively to the overall alpha value. As can be ascertained by examining the fifth column of the table (i.e., Alpha If Item Deleted), if items 9, 13, 14 and 15 were deleted, the alpha level would actually increase. Consider item 9. The mean value of this item (.9952) indicates that almost every single examinee in this sample answered this item correctly. The standard deviation (.0690) simply reflects the minute variability with which the examinees responded to this item. The negative corrected item-total correlation (-.0514) indicates that the manner in which the examinees responded to this item is negatively correlated with their total score. Table 4

Reliability: Part 1 (Aural) Listening Comprehension
(*n* = 210)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.9619	.1919	.3298	.5220
2	.9238	.2659	.2940	.5188
3	.7429	.4381	.2142	.5269
4	.9952	.0690	.1753	.5431
5	.8857	.3189	.2899	.5156
6	.9762	.1528	.3459	.5255
7	.4286	.4961	.2061	.5302
8	.8571	.3508	.1333	.5410
9	.9952	.0690	-.0514	.5505
10	.9619	.1919	.3795	.5173
11	.9905	.0974	.3444	.5333
12	.9524	.2135	.3187	.5207
13	.6095	.4890	.0416	.5688
14	.2667	.4433	.1185	.5478
15	.8571	.3508	.0933	.5478
16	.6952	.4614	.1857	.5338
17	.7571	.4298	.1797	.5342
18	.8571	.3508	.1468	.5387
19	.7476	.4354	.2331	.5227
20	.9381	.2416	.1605	.5368

Alpha= .5467

In other words, an examinee who did poorly on this section may have answered this item correctly and vice versa; one who did well on this section may have answered the item incorrectly. Lastly, if this item were deleted, the overall alpha level would increase from (.5467) to (.5505). Table 5 reflects a relatively low overall alpha level (.6773) for the Informal Words subtest. Items 2, 6 and 9 either contribute marginally or negatively to the alpha level, as indicated in the fifth column of the table.

Table 5

Reliability: Part 1 (Aural) Informal Words
(*n* = 210)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.9143	.2806	.3479	.6529
2	.8667	.3407	.1302	.6932
3	.9476	.2233	.3051	.6616
4	.8000	.4010	.3777	.6464
5	.9286	.2582	.4735	.6349
6	.9905	.0974	.3167	.6716
7	.8952	.3070	.5410	.6164
8	.8381	.3692	.3829	.6446
9	.7048	.4572	.2574	.6816
10	.8238	.3819	.4679	.6250

Alpha = .6773

Table 6, however, indicates a relatively higher level of internal consistency (.8374) for the Formal Equivalents subtest. Each of the ten items appear to be contributing to the overall internal consistency of this subtest. However, none of these three aural subtests demonstrated an acceptable coefficient alpha. Bachman (1990) suggests examining the role of the test format, the examinees' personal attributes, and random factors as potential sources which threaten internal consistency. From the perspective of test format, recall that this portion of the exam was intended to be administered in a language laboratory (Valdés, 1989). However, there is some evidence which indicates that this part of the Four Skills Exam is not always administered in such a facility (Navarrete & Benjamin, 1993). Moreover, each of these subtests require the examinee to read the test booklet and the Formal Equivalents subtest format even requires the examinee to spell the word correctly. It may be that these listening subtests are to some degree as much a reading and writing measure as they are a listening measure, a question addressed in the next chapter.

Table 6

Reliability: Part 1 (Aural) Formal Equivalents
(*n* = 210)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.7714	.4209	.4140	.8327
2	.6286	.4843	.5827	.8171
3	.5810	.4946	.4452	.8313
4	.6000	.4911	.5295	.8227
5	.1190	.3246	.4515	.8302
6	.3095	.4634	.6139	.8142
7	.2000	.4010	.5400	.8221
8	.3381	.4742	.5812	.8173
9	.3619	.4817	.5748	.8180
10	.4810	.5008	.5753	.8179

Alpha = .8374

Lastly, and as an example of a personal attribute, the Spanish language training the examinees received could be quite varied since the examinees were administered the exam at six different test sites which are likely to be affiliated with institutes of higher education where the examinees received their training. Moreover, the vast majority of the examinees appear to have Hispanic surnames, and as explained earlier, this segment of the population has not acquired its Spanish language skills in a supportive social milieu.

With respect to Table 7, which reflects the analyses related to the first oral passage common to each form of the Four Skills Exam, a seemingly different picture emerges. The overall coefficient alpha is somewhat high, (.8970). However, note the relatively uniform coefficient values across the categories of fluency, vocabulary, and appropriateness. There is little variability in terms of what each of these linguistic criteria contribute to the overall internal consistency of this oral protocol. In effect, the rater appears to be assigning a rating for fluency which determines the subsequent ratings for vocabulary and appropriateness. The result is a spuriously high alpha coefficient.

This phenomenon is referred to as the halo effect (Borg & Gall, 1979; Popham, 1990). According to Borg and Gall (1979), the rater forms an early impression of the person being observed and permits this initial impression to influence subsequent ratings. The authors also maintain that the halo effect is most likely to occur in the assessment of abstract qualities as opposed to specific behaviors.

There is additional testing facet evidence which supports this interpretation of the analyses. As noted in chapter two, the scoring procedures for the oral passages are not formally documented. More specifically, there are no benchmarks which anchor each of the possible ratings for the three different linguistic criteria rated. In other words, in the absence of a scoring manual, it is difficult to know what constitutes poor, weak, fair, good or very good fluency, vocabulary or appropriateness. The benchmarks exist in the minds of the two raters, and given that the linguistic criteria, especially fluency and appropriateness are abstract entities, it is not surprising that the raters fall victim to the halo effect.

Table 7

Reliability: Part 2 (Oral) Passage One
($n = 210$)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
Fluency	3.2286	1.2041	.9894	.7753
Vocabulary	3.2190	1.2256	.9913	.7747
Appropriateness	3.2286	1.2120	.9939	.7732

Alpha = .8970

With respect to Table 8, the overall alpha coefficient for the composition part of the test is also relatively high, (.8892). Nonetheless, note that the coefficients for communication, appropriateness and expression are nearly uniform across the four columns of values. The interpretation of these data is essentially the same as the argument set forth above for Oral Passage One. Once the scorer has assigned a rating for communication, this value influences the rating for appropriateness and expression. The end result is a spuriously high alpha coefficient.

Again, the spuriously high alpha coefficient is best explained as a consequence of the testing facets governing the Four Skills Exam. The halo effect overcoming the scorers would appear to rest once again on the absence of explicit scoring criteria, including benchmarks, for rating the compositions. Not having explicitly operationalized the relatively abstract linguistic criteria (i.e., communication, appropriateness, and expression) must be at the core of the matter.

It is important to recall that the rating of the composition consists of two steps, the first of which entails the three linguistic criteria reviewed above. If the examinee produces a composition that is on topic, consists of at least 150 words and generates an average overall rating of at least Fair (3.0) on these three criteria, the examinee's composition is then scored using a second set of criteria which is purely grammatical.

Table 8

Reliability: Part 4: Composition
(*n* = 210)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
Communication	2.7429	1.2102	.9304	.8210
Appropriateness	2.8000	1.2558	.9308	.8187
Expression	2.7381	1.2803	.9370	.8161
Final Rating	1.5238	1.8973	.8284	.8717

Alpha = .8892

If more than twenty errors are found within the first 150 words, the grading stops.

If less than twenty errors are found, each error is then assigned a numerical value ranging from 1 to 3, depending on the severity of the error. Once these errors have been identified and assigned their respective severity error value, these values are then summed to generate an error score. The rating scale for the final error score is as follows: 16 or less, Very Good; 17 to 22, Good; 23 to 30, Fair; 31 to 38, Weak; and 39 or more, Poor. The examinee must earn a Fair rating in order to pass this part of the exam.

An analyses of the inter-rater reliability of the scoring for composition errors was beyond the scope of this study. Nonetheless, at least the following observations can be made. First, and as previously stated, the scoring of the exam requires three scorers and only two are used. Suppose the two raters demonstrate a relatively high level of inter-rater reliability, for example (.85 to .90). Given the explicitness and concreteness of the scoring for composition errors, such a level is not out of the question.

The critical issue concerns the reliability of the error score for making a valid judgment about the writing ability of the examinee. The point is that while the scoring process may be consistent, the interpretation of the error score renders the reliability of the judgments questionable. A composition containing twenty one-point grammatical errors entailing punctuation, capitalization, accentuation, spelling, etc. does not seem appropriate for the purpose of professional educational licensure.

Based on these analyses the Common Parts of the Four Skills Exam appear to be lacking substantive reliability. Only the Formal Equivalents subtest which forms part of the aural portion of the test demonstrated what might be considered adequate internal consistency (.8374). The low internal consistency of the other two aural subtests appear to be linked to the test method facets. Similarly, the spuriously high alpha coefficients generated for Oral Passage One and the Composition are likely due to test method facets, mainly the absence of explicit scoring criteria and benchmarks.

Uncommon Parts: Form A

Table 9 which provides the coefficients of internal consistency for the Dictation subtest, reveals a moderate overall alpha coefficient, (.8497). As noted in Chapter 2, the examinee must listen to an audio taped stimuli and fill in the blank with the missing word(s). In order to answer the item correctly, the examinee must also spell the word(s) correctly. Items 3, 10, 16 and 19 appear to contributing negligibly or negatively to the overall internal consistency of this subtest as can be surmised through an examination of the Alpha If Item Deleted column. Nearly all of the examinees answer item 3 correctly; items 19 and 16 follow in terms of their facility.

Item 10 does not appear to be answered consistently correct by either those examinees that score well or poorly on this subtest. Exactly half of the examinees answer this item correctly as the mean value reveals (.5000). The standard deviation for this item, (.5018), is also the largest among the twenty test items. The corrected item-total correlation is accordingly relatively low (.2909).

It is somewhat difficult to explain this moderate alpha coefficient vis a vis the previously examined aural subtests. Recall that the formal equivalents subtest also yielded a moderate alpha coefficient (.8374). On the other hand, the two remaining subtests yielded low alpha coefficients, (.5467) for the Listening Comprehension subtest and (.6773) for the Informal

Words subtest. All four subtests are tape-mediated but apparently the internal consistency of each subtest is impacted differently by the lack of appropriate language laboratory facilities for their administration.

Table 9

Reliability: Part 1 (Aural) Dictation
(*n* = 140)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.8786	.3278	.4548	.8426
2	.8857	.3193	.2926	.8480
3	.9929	.0845	.2035	.8507
4	.8143	.3903	.5260	.8392
5	.7286	.4463	.3650	.8461
6	.8286	.3782	.4145	.8437
7	.8071	.3960	.4135.	.8437
8	.8143	.3903	.5359	.8388
9	.6357	.4830	.4062	.8446
10	.5000	.5018	.2909	.8507
11	.5857	.4944	.5820	.8357
12	.1429	.3512	.4106	.8439
13	.6286	.4849	.6568	.8318
14	.6286	.4849	.5649	.8366
15	.6143	.4885	.4444	.8428
16	.8357	.3719	.2658	.8493
17	.7786	.4167	.5927	.8360
18	.7429	.4386	.5618	.8371
19	.8786	.3278	.2689	.8488
20	.1286	.3359	.3360	.8466

Alpha = .8497

Perhaps this finding is best explained in the following manner. The Listening Comprehension and Informal Equivalents subtests both use the same test format, multiple-choice. In contrast, the Dictation and Formal Equivalents use a fill in the blank format. Perhaps the inconsistent use or availability of language laboratory facilities coupled with a multiple-choice format create greater hardships for the examinees.

The multiple-choice formats also entail more reading than the fill in the blank format. On the other hand, the fill in the blank subtests require a correctly written one word or short phrase response. A more conclusive explanation can only be achieved once the administration of this portion of the exam is more fully standardized.

Table 10 and 11 contain the alpha coefficients for Oral Passage Two and Oral Passage Three, respectively. Based on the alpha level (.9141) reported for Oral Passage Two in Table 10, one may be inclined to surmise that this portion of the exam has a moderately high internal consistency. However, upon closer inspection of the means and correlations contained within the table, there is very little variability among the coefficients. This uniformity may well be signaling the influence of the halo effect described earlier with respect to Oral Passage One.

Table 10

Reliability: Part 2 (Oral) Passage Two
(*n* = 140)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
Fluency	2.9571	1.5167	.9820	.8200
Vocabulary	2.9000	1.5561	.9766	.8231
Appropriateness	2.9429	1.5351	.9888	.8173

Alpha = .9141

Perhaps the most compelling evidence which substantiates the presence of the halo effect is contained within the Corrected Item-Total Correlation column. Again, Borg and Gall (1979) describe this correlation as the strength of relationship between the item score with the total score. There is a near perfect correlation (1.000) between the three linguistic criteria (i.e., fluency, vocabulary, and appropriateness) and the overall alpha. According to these data, the raters could essentially base their judgments on any one of the three rated criteria; rating all three criteria is redundant.

Table 11

Reliability: Part 2 (Oral) Passage Three
(*n* = 140)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
Fluency	3.3857	1.0967	.9470	.7621
Vocabulary	3.1643	1.2672	.9332	.7736
Appropriateness	3.3214	1.1460	.9767	.7464

Alpha = .8827

Essentially the same scenario unfolds for Oral Passage Three summarized by Table 11. The correlations between the three linguistic criteria and the total score is again quite high, though slightly lower than the correlations evidenced for the other two oral passages.

The interpretation of this phenomenon is again linked to the test method facets of the exam: the use of only two scorers when three are required, the lack of documentation substantiating inter-rater reliability, the absence of a scoring manual with explicit benchmarks for fluency, vocabulary and appropriateness, and the administration of this portion of the exam without the recommended language laboratory facilities.

Table 12 contains the coefficients for the Orthography: Accents subtest. The overall alpha coefficient is moderately low (.8194). There are at least eight items (i.e., items 1, 5, 7, 11, 14, 16, 17 and 19) which contribute minimally to the overall internal

Table 12

Reliability: Part 3 (Reading) Orthography: Accents
(*n* = 140)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.6214	.4868	.2505	.8194
2	.6714	.4714	.4811	.8066
3	.6071	.4901	.4671	.8073
4	.7571	.4303	.4884	.8067
5	.7143	.4534	.2868	.8169
6	.8071	.3960	.4479	.8091
7	.6000	.4917	.2537	.8193
8	.7071	.4567	.4203	.8100
9	.4929	.5017	.4165	.8102
10	.7286	.4463	.3930	.8115
11	.8500	.3584	.3043	.8156
12	.7500	.4346	.4775	.8072
13	.4929	.5017	.6446	.7966
14	.7786	.4167	.2464	.8185
15	.8357	.3719	.4534	.8092
16	.8429	.3652	.3102	.8153
17	.6429	.4809	.3240	.8153
18	.6571	.4764	.4107	.8105
19	.7643	.4260	.2665	.8176
20	.7429	.4386	.4371	.8092

Alpha = .8194

consistency of this subtest as the Alpha If Item Deleted data reveal. The corrected item-total correlation for these eight items range between (.2505) and (.3240) which provides further evidence that these items merit closer inspection.

The Orthography: Accents subtest, like the Dictation and Formal Equivalents aural subtests, appears to hold its own. That is, there does not appear to be any one particular source of error which might be detracting from its reliability. Since this multiple-choice subtest is administered in a straight forward manner, there are fewer sources of potential error. Unlike the aural and oral sections of the exam which require language laboratory facilities, all the examinee needs for this portion of the test is essentially a legible test booklet.

Table 13, which reflects data for the Orthography: Spelling subtest, yields another moderate alpha coefficient (.8045). The items which appear to be contributing little to the overall internal consistency of this subtest are items 2, 5, 7, 9, 15, and 18. This observation is supported by the coefficient values reflected for these items in the Corrected Item-Total Correlation and Alpha If Item Deleted columns. The range for the item-total correlation for these items is between (.1960) and (.2834). Items 5 and 15 appear to be relatively easy for the examinees as the respective mean values for these items reveal. Again, the internal consistency of this subtest may be attributable to the straight-forwardness of the testing format and task, spelling.

The test items examined in Table 14 which constitute the Identifying Concepts subtest reveal a relatively low level of internal consistency (.7732). Interestingly, the correlations in this subtest appear to cluster together in two parts. As reflected in the Alpha If Item Deleted column, Items 1 through 8 contribute weakly to the overall alpha coefficient: in contrast, items 9 through 16 contribute in a positive manner to the internal consistency of this test. The coefficients generated for the Corrected Item-Total Corre-

Table 13

Reliability: Part 3 (Reading) Orthography: Spelling
(*n* = 140)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.6714	.4714	.4021	.7944
2	.7429	.4386	.1960	.8067
3	.6143	.4885	.4776	.7893
4	.8929	.3104	.3479	.7979
5	.9571	.2033	.1436	.8053
6	.4571	.4999	.4998	.7877
7	.7857	.4118	.2834	.8011
8	.4857	.5016	.4741	.7895
9	.8786	.3278	.2322	.8029
10	.8643	.3437	.3574	.7972
11	.5071	.5017	.3858	.7958
12	.5857	.4944	.3272	.7997
13	.8071	.3960	.3428	.7978
14	.8857	.3193	.3959	.7957
15	.9786	.1453	.2185	.8038
16	.7357	.4425	.4388	.7921
17	.8000	.4014	.5225	.7877
18	.8857	.3193	.2755	.8010
19	.5429	.4999	.5536	.7839
20	.7429	.4386	.4264	.7929

Alpha = .8045

tion column conform to this pattern as well. The range of weak correlation values for the first eight test items range from (-.0272) to (.2317); on the other hand, the correlations for items 9 through 16 are much stronger, (.4271) to (.7092). In effect, the two sets of clustered items may be acting as if they constituted two distinct subtests.

This clustering effect is probably due to the fact that this sub-section contains two different testing formats. For items 1 through 8 the examinee must read a short passage or statement and then answer a multiple-choice question. For the remaining items, the examinee must read a passage and then order four statements in the sequence they appear in

Table 14

Reliability: Part 3 (Reading) Identifying Concepts
(*n* = 140)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.9714	.1672	.0782	.7774
2	.9429	.2329	.1860	.7736
3	.9357	.2461	.1945	.7734
4	.9929	.0845	-.0272	.7783
5	.8429	.3652	.1889	.7795
6	.9857	.1191	.0937	.7758
7	.9357	.2461	.0371	.7833
8	.9714	.1672	.2317	.7708
9	.9143	.2809	.4271	.7570
10	.8071	.3960	.5003	.7486
11	.7571	.4303	.5143	.7472
12	.7000	.4599	.7092	.7212
13	.8571	.3512	.5496	.7441
14	.9000	.3011	.5647	.7450
15	.9143	.2809	.4901	.7521
16	.8286	.3782	.6472	.7325

Alpha = .7732

the passage. The fundamental difference between the two test formats is the length of the reading passages. The first four items stem from short paragraphs, while the following four items stem from four separate statements or sentences. The remaining eight items are based on two reading passages each of which consists of approximately 150 words. In effect, the testing format used for the first eight items should be reexamined in order to create a more reliable reading measure.

Correlation data for the final reading subtest, Words in Context, are summarized in Table 15. The overall alpha this subtest is low (.6739). While there are only four items which

Table 15

Reliability: Part 3 (Reading) Words in Context
(*n* = 140)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.5714	.4966	.5059	.6107
2	.6071	.4901	.3868	.6394
3	.8429	.3652	.1631	.6786
4	.9357	.2461	.1998	.6709
5	.7286	.4463	.4838	.6186
6	.5786	.4956	.3893	.6389
7	.7429	.4386	.3684	.6434
8	.9786	.1453	.1960	.6735
9	.8214	.3844	.1087	.6886
10	.5643	.4976	.4889	.6149

Alpha = .6739

detract from the internal consistency of this subtest, there are only ten items in the entire subtest. Items 3, 4, 8 and 9 are highly suspect in terms of their contribution to the reliability of this subsection. Interestingly, these four items are most often answered correctly by the examinees as the Mean column reveals.

With respect to the potential source of error in this sub-section, the testing format seems suspect. For this subtest the examinee must read a short passage with a missing word or words. Then the examinee must select the correct answer from among three options. Perhaps the correct option is too obvious or the accompanying distractors too easily discarded. The format itself, a contrived cloze procedure, does not appear to be problematic since the remaining six test items appear to be functioning efficiently.

Uncommon Parts: Form B

The alpha coefficient for the Dictation subtest presented in Table 16 indicates a moderate degree of internal consistency (.8649). Items 2 and 16 appear to be the items which are contributing least to the reliability of this subtest. Similarly, items 5 and 3 contribute negligibly. The negative corrected item-total correlation for item 2 indicates that those examinees who do well on this subtest do most poorly on this item. Interestingly, this item is the easiest for the examinees as the mean value of this item indicates (.9833).

In terms of which source of the testing process might be detracting from the internal consistency of this subtest, it appears safe to assume that source is at the item level. Apparently, the lack of prescribed language laboratory facilities for the administration of this subtest does not affect its reliability as much as it may be affecting some of the other aural language subtests. Again, the examinee needs only hear the sentence, and then supply the missing word ensuring for correct spelling. In brief, the subtest format is straight-forward.

A closer examination of the four identified items may well enhance the internal consistency of this subtest.

Tables 17 and 18 indicate that Oral Passage Two and Three are again relatively high in terms of their internal consistency, (.9157) and (.8934) respectively. Nonetheless, and as previously discussed with respect to the oral passages reviewed to this point, there is a high degree of uniformity in terms of the coefficient values reported in

Table 16

Reliability: Part I (Aural) Dictation
(*n* = 60)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.9167	.2787	.5662	.8567
2	.9833	.1291	-.0104	.8682
3	.9333	.2515	.3081	.8634
4	.7167	.4544	.5775	.8537
5	.9500	.2198	.1126	.8674
6	.8667	.3428	.4940	.8577
7	.1500	.3601	.3887	.8611
8	.6667	.4754	.6881	.8485
9	.7000	.4621	.5604	.8545
10	.8167	.3902	.5126	.8567
11	.8833	.3237	.6611	.8528
12	.6333	.4860	.4608	.8590
13	.6000	.4940	.7064	.8474
14	.6167	.4903	.4055	.8616
15	.6333	.4860	.6510	.8502
16	.3833	.4903	.0993	.8750
17	.6333	.4860	.4054	.8615
18	.8667	.3428	.4558	.8590
19	.9000	.3025	.4907	.8583
20	.8000	.4034	.5075	.8568

Alpha = .8649

both tables. Again, the most compelling evidence stems from the Corrected Item-Total Correlation columns. Both tables reflect coefficients which are nearly perfectly correlated with the overall alpha. The potential reasons underlying these spuriously high alpha coefficients have already been detailed with respect to the halo effect. On the positive side, at least the scorers are being consistent from passage to passage and form to form.

Table 17

Reliability: Part 2 (Oral) Passage Two
(*n* = 60)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
Fluency	2.5333	1.8362	.9825	.8238
Vocabulary	2.4500	1.8266	.9819	.8239
Appropriateness	2.5000	1.8273	.9953	.8181

Alpha = .9157

Table 18

Reliability: Part 2 (Oral) Passage Three
(*n* = 60)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
Fluency	2.9667	1.4840	.9793	.7714
Vocabulary	2.8667	1.5123	.9787	.7721
Appropriateness	2.9333	1.4714	.9838	.7692

Alpha = .8934

The overall alpha coefficient for Orthography: Accents contained in Table 19 is less than moderate (.7489). There are six items which appear to be weakening the internal consistency of this subtest, items 7, 8, 13, 14, 15, and 17. The corrected item-total correlations range between (.0300) and (.2404). Given the fact that these multiple-choice items do not appear to be difficult for the examinees, the most probable source of error may be the accompanying distractors.

Table 19

Reliability: Part 3 (Reading) Orthography: Accents
(*n* = 60)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.8667	.3428	.3329	.7381
2	.8500	.3601	.2965	.7403
3	.7833	.4155	.2659	.7425
4	.5833	.4972	.3993	.7314
5	.6500	.4810	.5627	.7163
6	.6000	.4940	.3448	.7364
7	.9167	.2787	.1050	.7508
8	.8167	.3902	.1844	.7481
9	.6833	.4691	.3733	.7339
10	.8500	.3601	.2820	.7412
11	.8333	.3758	.4059	.7326
12	.6833	.4691	.5774	.7154
13	.8667	.3428	.1820	.7475
14	.7000	.4621	.0300	.7623
15	.6333	.4860	.1312	.7551
16	.8500	.3601	.4142	.7324
17	.7333	.4459	.2404	.7449
18	.8333	.3758	.3208	.7385
19	.7667	.4265	.4712	.7263
20	.8000	.4034	.2905	.7406

Alpha = .7489

As Table 20 indicates, the overall alpha for the Orthography: Spelling subtest is (.8784), a moderately high reliability coefficient. Namely items 12, 15 and 18 are not pulling their weight in this subtest. Items 15 and 18 appear to be particularly easy for the examinees as the mean values for these items imply.

A closer examination of items 12, 15 and 18 would potentially enhance the reliability of this subtest. Recall the test format for this section of the exam is one in

Table 20

Reliability: Part 3 (Reading) Orthography: Spelling
(*n* = 60)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.8333	.3758	.5926	.8693
2	.8500	.3601	.5517	.8707
3	.8333	.3758	.5238	.8715
4	.7667	.4265	.6007	.8685
5	.8833	.3237	.5078	.8724
6	.8833	.3237	.3408	.8769
7	.8000	.4034	.4170	.8750
8	.7333	.4459	.5208	.8714
9	.9000	.3025	.5786	.8708
10	.3500	.4810	.5365	.8710
11	.7833	.4155	.6304	.8675
12	.5667	.4997	.2234	.8839
13	.5333	.5031	.4433	.8751
14	.8000	.4034	.5121	.8717
15	.9333	.2515	.1606	.8805
16	.7000	.4621	.7315	.8630
17	.4333	.4997	.5851	.8690
18	.9833	.1291	.1900	.8796
19	.8500	.3601	.4573	.8736
20	.8333	.3758	.5238	.8715

Alpha = .8784

which the examinee is required to read a short sentence and to detect a misspelled word. The examinee must then spell the word correctly. Perhaps the misspelled words are not as difficult for the examinees to identify and to correctly spell as the test authors might have assumed.

Table 21 reflects coefficient patterns similar to those reviewed earlier for the same subtest in Form A. Five items are detracting from the internal consistency of the subtest, and are embedded within the first eight test items, specifically items 1, 3, 5, 6, and 7. Moreover,

Table 21

Reliability: Part 3 (Reading) Identifying Concepts
(*n* = 60)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	0.9167	.2787	.0071	.7790
2	0.9667	.1810	.4419	.7481
3	1.0000	.0000	-----	-----
4	0.9667	.1810	.6323	.7372
5	0.9833	.1291	.0233	.7685
6	0.8333	.3758	.1374	.7760
7	1.0000	.0000	.----	-----
8	0.9500	.2198	.5397	.7387
9	0.9167	.2787	.2113	.7624
10	0.7333	.4459	.4795	.7375
11	0.7500	.4367	.4917	.7355
12	0.9667	.1810	.3031	.7558
13	0.9333	.2515	.5605	.7342
14	0.8167	.3902	.5175	.7317
15	0.8500	.3601	.5752	.7254
16	0.7833	.4155	.5235	.7308

Alpha = .7621

items 3 and 7 contribute nothing to the reliability of this subtest. Only item 9, which forms part of a different testing format, is not part of the cluster which contributes the most to the internal consistency of this subtest. As noted previously with respect to the Identifying Concepts subtest in Form A, the testing format for items 1 through 8 should be carefully reviewed. The format seems to be one that does not require the examinees to engage in a reading process that challenges their comprehension. The reading passages for the first four items are brief in comparison to the reading passages used for items 9 through 16. Moreover, items 5 through 8 are based on four sentence length statements. In addition, if the cloze procedure is going to be used in conjunction with a multiple-choice format, a closer examination of the distractors is warranted especially for the first four items.

Table 22 reveals that the Words in Context subtest also falls short of adequate

Table 22

Reliability: Part 3 (Reading) Words in Context
(*n* = 60)

Item	Mean	Std Dev	Corrected Item-Total Correlation	Alpha If Item Deleted
1	.9333	.2515	.0961	.6510
2	.9667	.1810	.2561	.6350
3	.7833	.4155	.3727	.6053
4	.6500	.4810	.3593	.6077
5	.7500	.4367	.5053	.5732
6	.6000	.4940	.4517	.5831
7	.2833	.4544	.4581	.5834
8	.5667	.4997	.1781	.6539
9	.6833	.4691	.2262	.6397
10	.9167	.2787	.1915	.6394

Alpha = .6439

internal consistency. The overall alpha for this section is (.6439). Items 1 and 8 are clearly contributing little to the internal consistency of this measure. The Alpha If Item Deleted coefficients and the Corrected Item-Total Correlation for these two items bear this out. While items 9 and 10 do contribute positively to the internal consistency of this subtest, their impact is only marginal. This observation is especially supported by the weak Corrected Item-Total Correlations for items 9 and 10, (.2262) and (.1915), respectfully. While there appear to be at least four items which contribute marginally to the overall internal consistency of this measure, there is something particularly troublesome about the test format. The examinees read a paragraph from which one word has been omitted. The examinee must then select the correct word from one of four options. In effect, for some items there is no need to read the entire paragraph to answer the item; the examinee is may actually only have to read the sentence from which the word has been deleted. In short, careful consideration should be given to the use of this test format and a closer examination of the accompanying distractors is needed.

Summary

Table 23 summarizes all of the alpha coefficients for the different subtests across Forms A and B of the Four Skills Exam. At the beginning of this chapter, a decision was made to establish the acceptable level of reliability at the (.90) level. This standard is warranted since the Four Skills Exam is a high-stakes exam which affects teachers' and students' lives.

The three alpha coefficients generated for the Common Parts of the aural section of the Four Skills Exam reveal that the Formal Equivalents subtest is nearest to demonstrating acceptable internal consistency, (.8374). Both forms of the Dictation subtest also demonstrated a moderate level of internal consistency, (.8497) for Form A and (.8649) for Form B. In fact, only two of the aural language subtests have the potential of achieving an acceptable level of reliability. This potential resides in a reexamination of the test items contributing marginally to the internal consistency of the subtests.

Table 23

Summary of Reliability Across Test Forms

Test Part & Subtest	Common Parts (n = 210)	Form A (n = 140)	Form B (n = 60)
(Aural) Listening Comprehension	.5467		
(Aural) Dictation		.8497	.8649
(Aural) Informal Equivalents	.6773		
(Aural) Formal Equivalents	.8374		
(Oral) Passage One	.8970		
(Oral) Passage Two		.9141	.9157
(Oral) Passage Three		.8827	.8934
(Reading) Orthography: Accents		.8194	.7489
(Reading) Orthography: Spelling		.8045	.8784
(Reading) Identifying Concepts		.7732	.7621
(Reading) Words in Context		.6739	.6439
(Writing) Composition	.8892		

More importantly and given that the four aural language subtests contain a total of sixty items, and the Dictation and Formal Equivalents subtests account for thirty of these items, only half of the items or subtests in the aural part of the test are potentially reliable enough on which to base a decision. None of the aural subtests are internally consistent enough if one adheres strictly to the (.90) cutoff previously established. As a result, the aural part of the Four Skills Exam does not appear to be reliable enough to support a valid decision regarding the aural language ability of an individual examinee. With respect to the oral part of the Four Skills Exam, there appears to be ample evidence which supports the presence of a halo effect (Borg and Gall, 1979; Popham, 1990). The same case also appears to apply to the written part of the test as measured through the Composition. Consequently, the alpha coefficients reported in Table 23 are spuriously high.

In terms of the reading comprehension part of the exam, none of the four subtests within Form A of the test demonstrated an adequate level of internal consistency. Of the four subtests comprising Form A, the Orthography: Accents and Orthography: Spelling subtests were the most reliable. Nonetheless, making a valid judgment about the reading ability of an examinee based on the examinee's reading score would appear to prove problematic. This is especially true for those examinee's taking Form B of the exam since only the Orthography: Spelling subtest demonstrated near sufficient internal consistency.

As indicated at the beginning of this chapter, reliability analyses could not be conducted on Form C of the exam since there were so few cases (*n* = 10) on which to base the analyses. Interestingly, nothing is known about the reliability of Form C. Neither the analyses conducted by Cárdenas (1981) or Young, et al (1986) included this form of the test. Moreover, the fact that Form C of this exam was used only ten times over 1991-92 raises a separate but related issue. Clearly, the different forms of the test are not being systematically rotated, a strategy which helps maintain the reliability of the equivalent forms of a test.

With respect to the evaluative framework advocated by Messick (1989), quite a lot can be said. Since the four parts of the Four Skills Exam are weak in terms of their reliability, the construct validity of the measure is questionable. Recall that a test that is not reliable cannot be valid (Bachman, 1990). Furthermore, making valid judgments about prospective bilingual education teachers' Spanish language proficiency in NM based on such test scores would clearly prove problematic. At best, the Four Skills Exam is able to generate moderately reliable measures of the spelling and orthographic abilities of the examinees. The social consequences of using this test for endorsing teachers are not favorable. The primary objective of this test is to help protect the rights of students in need of Spanish language instruction. Given the evidence presented in this chapter, the Four Skills Exam falls short of fulfilling the social function for which it was intended.

CHAPTER 5

CORRELATIONS & EQUIVALENCY OF FORMS

In this chapter, the correlations among the various subtests, test parts and common and uncommon parts of the Four Skills Exam are presented and examined. These analyses provide one piece of construct related evidence which is needed in order to assist in making an overall evaluative judgment concerning the validity of the instrument in question. One would expect to find particular patterns of correlations among the different subtests, subtests within test parts, and so forth. For example, one would expect for the four aural subtests to be at least moderately correlated with one another since each subtest purports to measure the same construct, primarily the prospective bilingual education teacher's ability to auditorily comprehend spoken Spanish. As a contrasting example, one would not expect to find a strong correlation between the oral and reading comprehension parts of the test since the two parts are designed to measure different, though related, constructs.

The statistical procedure used to generate these data was the product-moment correlation coefficient (r) commonly referred to as the Pearson correlation coefficient.

Correlations between two variables (e.g., two subtests) can range from -1.0 to 1.0. A coefficient with a negative value indicates that there is an inverse relationship between the two variables. Negative correlations should not be found since all the variables are related to the more general construct, in this case language proficiency. Similarly, a correlation of zero indicates that there is no relationship between the two variables. A positive correlation, which can range from a value greater than zero up to 1.0, indicates that there is a positive relationship between the two variables. As alluded to in the above paragraph, the nature of this positive relationship should vary as a function of the constructs being correlated.

In actuality, one would expect to find low correlations among subtests which purport to measure very different language abilities (e.g., oral proficiency versus reading comprehension) and moderate correlations between measures which tap overlapping language abilities such as aural comprehension and oral production. High correlations are likely to be found between two test parts which measure the same construct. Bachman (1990) maintains that there are essentially three possible interpretations one can infer regarding a given correlation. That is, the correlation is a function of the: (1) trait being measured; (2) the test format; and (3) a combination of the trait and test format. The inferences made by this author are guided by the insight offered by Bachman.

Correlations: Form A

Table 24 contains the first set of correlations between the common parts of the test and the common parts of the test to the test parts (i.e., Parts 1 through 4). There appears to be a moderate positive correlation (.4499) between the performance of the examinees on the Informal Words (IWORDS) subtest and the Listening Comprehension

(LCOMP) subtest. This moderate correlation is most likely due to the fact that both subtests purport to measure the same construct, listening comprehension. Moreover, both subtests utilize a multiple-choice format. In contrast, a weak correlation (.1884) was found between the Informal words subtest and the Formal Equivalents (FEQUIV) subtest. This is most likely due to the fact that the former employs a multiple-choice format while the latter test format requires the examinee to write and correctly spell the test answer.

Similarly weak correlations exist between Oral Passage One (OPAS1) and the Listening Comprehension subtest (.1909) and the Informal Words subtest (.0579). One would have expected to find a somewhat stronger correlation between aural and oral measures. Again, recall that the Informal Words subtest requires reading on the part of the examinee. A stronger correlation (.4028) is in fact found between the Formal Equivalents subtest and Oral Passage One, however. Perhaps this is due to the fact that the examinee is rated on vocabulary in Oral Passage One and in a like manner the correct spelling of a word on the Formal Equivalents subtest.

Relatively weak positive correlations were also found between the four reading subtests and the Listening Comprehension and Informal Words subtest. These correlations ranged from (.0187) to (.3050). However, relatively higher correlations were found between the four reading subtests and the Formal Equivalents subtest. Again, recall that the Formal Equivalents subtest, while an aural measure, also requires correct spelling. In fact, the strongest correlation (.6925) was found between the Formal Equivalents subtest and the Orthography: Spelling subtest. Both essentially require the same language skill, correct spelling.

With respect to the correlations between the three aural subtests to the four different parts of the test, the most salient relationships again appear to exist between the performance of the examinee on the Formal Equivalents subtest and the four parts of the exam. A relatively strong correlation (.8641) exists between the performance of the examinee on the Formal Equivalents subtest and their performance on Part 1 (Aural).

Table 24

Correlations Between Aural & Oral Common Parts and Subtests & Common Parts

	LCOMP	IWORDS	FEQUIV	OPAS1
Listening Comprehension	1.0000			
Informal Words	0.4499**	1.0000		
Formal Equivalents	0.3316**	0.1884**	1.0000	
Oral Passage One	0.1909**	0.0579	0.4028**	1.0000
Orthography: Accents	0.1341	0.0187	0.6345**	0.3002**
Orthography: Spelling	0.2454**	0.0893	0.6925**	0.3548**
Identifying Concepts	0.1557*	0.0852	0.3654**	0.2404**
Words in Context	0.3050**	0.1047	0.6324**	0.3468**
Part 1 (Aural)	0.6071**	0.4435**	0.8641**	0.4063**
Part 2 (Oral)	0.1583*	0.0592	0.4451**	0.7464**
Part 3 (Reading)	0.2491**	0.0835	0.7465**	0.3827**
Part 4 (Composition)	0.1324	0.0892	0.6154**	0.3256**
Common Parts	0.2903**	0.1918**	0.7191**	0.4854**

Also note the moderate to moderately high correlations between the Formal Equivalents subtest with Part 3 (Reading) and Part 4 (Composition), (.7465) and (.6154). Again, the problem appears to be that the Formal Equivalents subtest, which is intended to generate a measure of the examinee's aural comprehension, generates a measure that correlates more with reading and writing abilities as opposed to oral abilities. In fact, the correlations between Oral Passage One and Part 3 (Reading) and Part 4 (Composition) are more clearly defined. Lastly, and with respect to the correlations between the four subtests in question and the Common Parts variable, again the strongest correlation (.7191) exists between the Formal Equivalents subtest and the Common Parts. This can probably be attributed to the strong literacy orientation of the subtests which collectively comprise the Common Parts variable. Only the Listening Comprehension subtest and Oral Passage One do not require reading or writing; the remaining six subtests require either reading or writing or a combination of these two skills.

Table 25 contains the correlation coefficients between the four reading subtests and the remaining common parts and parts of the Four Skills Exam. There is clearly more moderate correlations between the four different reading subtests than was evidence above between the three aural subtests. None of the correlations among the reading subtests were especially low or high as they ranged between (.3803) and (.6905). Not surprisingly, the two orthography subtests share a moderate correlation of (.6905).

With regard to the correlations between the reading subtests and the four test parts, the most salient pattern is found between three of the reading subtests and Part 1 (Aural). Note the moderate correlation (.7497) between Orthography: Spelling (OSPEL) and Part 1 (Aural). Again, the aural subtests appear to be tapping the literacy skills of the examinee as three of the four aural subtests require either some reading or writing. A clearer contrast is found between the four reading subtests and Part 2 (Oral) of the exam. It is also somewhat unfortunate that the two reading subtests, Identifying Concepts (IDCON) and Words in Context (WCTEX) which correlate most highly with Part 3 (Reading) require the least of amount of reading from the examinee. That is, the Orthography: Accents (OACC) and Orthography: Spelling subtests are just what their heading suggests, measures of orthography or writing. Moreover, of the four reading subtests, it is in fact the Orthography: Accents and Orthography: Spelling which correlate most readily with Part 4 (Composition), (.5739) and (.5417).

In terms of the correlations found between the four reading subtests and the Common Parts of the exam, there appears to be a pattern of moderate positive correlations ranging from (.3067) to (.6052). This pattern is most likely attributed to the strong literacy orientation underlying the Common Parts variable explained above.

Table 25

Correlations Between Reading Comprehension Subtests, Test Parts & Common Parts

	OACC	OSPEL	IDCON	WCTEX
Orthography: Accents	1.0000			
Orthography: Spelling	0.6905**	1.0000		
Identifying Concepts	0.3803**	0.4885**	1.0000	
Words in Context	0.4236**	0.6250**	0.5061**	1.0000
Part 1 (Aural)	0.6003**	0.7497**	0.3948**	0.6151**
Part 2 (Oral)	0.3498**	0.3624**	0.1851**	0.3887**
Part 3 (Reading)	0.8585**	0.9195**	0.6321**	0.7325**
Part 4 (Composition)	0.5739**	0.5417**	0.2671	0.4465**
Common Parts	0.6036**	0.6052**	0.3067**	0.5141**

The final table related to the Common Parts of the Four Skills Exam, Table 26, offers correlation data at the level of the four test parts. It is interesting to note that Part 1 (Aural) correlated more highly with the Reading, Composition (COMPOS), and Common Parts (CMNPRTS) variable than with the Oral portion of the test. This pattern does not lend validity to the Aural part of this test. That is, the aural section appears to be measuring

Table 26

Correlations Between Test Parts & Common Parts Variable

	AURAL	ORAL	READING	COMPOS	CMNPRTS
Part 1	1.0000				
Part 2	0.4283**	1.0000			
Part 3	0.7568**	0.4075**	1.0000		
Part 4	0.5676**	0.3844**	0.6074**	1.0000	
Common Parts	0.7016**	0.4994**	0.6643**	0.9708**	1.0000

a language ability which more readily correlates with measures of literacy. The correlations between Part 2 (Oral) and the other four variables appear to be quite stable. In short, the oral portion of the test appears to be fulfilling its purpose, rendering a measure of oral proficiency which is not confounded by literacy.

Correlation Coefficients (Form A)

Table 27 summarizes the correlations between the four aural subtests and the entire battery of subtests, test parts, the common parts variable, uncommon parts variable and the total test score variable for Form A. With regard to the manner in which the aural subtests correlate with one another, two observations are warranted. First, it is interesting to note the moderately high positive correlation (.7450) between the Dictation (DICT) and Formal Equivalents (FEQUIV) subtest. This comes as no real big surprise since both subtests require the examinee to correctly spell words. In contrast, the Informal Words (IWORDS) subtest shares weak correlations with both the Dictation and Formal Equivalents subtests. Perhaps this is due to the fact that the Informal Words subtest is designed to measure the New

Mexican regional, lexical knowledge of the examinee. The low correlations may also be attributed to the spelling requirement inherent to the Dictation and Formal Equivalents subtests.

Table 27

Correlations Between Aural Subtests & All Remaining Test Variables

	LCOMP	DICT	IWORDS	FEQUIV
Listening Comprehension	1.0000			
Dictation	0.2252**	1.0000		
Informal Words	0.4829**	0.0825	1.0000	
Formal Equivalents	0.3067**	0.7450**	0.1861*	1.0000
Oral Passage One	0.2068*	0.4342**	0.0756	0.4668**
Oral Passage Two	0.1431	0.3163**	0.0505	0.3773**
Oral Passage Three	0.1045	0.4112**	0.1322	0.4845**
Orthography: Accents	0.0948	0.6676**	-.0214	0.6166**
Orthography: Spelling	0.2787**	0.8405**	0.0682	0.7133**
Identifying Concepts	0.1257	0.3759**	0.0292	0.3562**
Words in Context	0.3104**	0.5511**	0.0471	0.6076**
Part 1 (Aural)	0.6047**	0.8536**	0.4476**	0.8581**
Part 2 (Oral)	0.1903*	0.4798**	0.1038	0.5515**
Part 3 (Reading)	0.2398**	0.7993**	0.0358	0.7357**
Part 4 (Composition)	0.0321	0.5772**	0.1224	0.6018**
Common Parts	0.2103*	0.6503**	0.2332**	0.7218**
Uncommon Parts	0.2323**	0.8222**	0.0692	0.7706**
Total	0.2379**	0.7755**	0.1777*	0.7971**

With respect to the correlations between the listening subtests and the three oral passages, both the Listening Comprehension (LCOMP) and Informal Words subtests correlate weakly with the three oral passages. The correlations for this comparison range from a low of (.0505) to (.2068). The Dictation and Formal Equivalents subtests correlate more readily with the three oral passages, ranging between (.3163) and (.4845). This is somewhat puzzling since the Dictation and Formal Equivalents subtests require correct spelling, an ability which is not necessarily correlated with oral proficiency. In short, one would expect to find stronger correlations between the two listening subtests that do not require writing to correlate more strongly with the oral passages.

Again, a similar pattern emerges when one examines the manner in which the Listening Comprehension and Informal Words subtests correlate with the four reading subtests. The weakest correlations are also evidenced in this set of comparisons. The presence of the negative correlation (-.0214) between the Informal Words subtest and the Orthography: Accents subtests indicates that the two subtests are measuring two constructs which are not necessarily language related which of course should not be the case.

As previously mentioned, the Dictation and Formal Equivalents aural subtests correlate moderately with all four reading subtests. In effect, the Dictation and Formal Equivalents subtests are more a measure of literacy than listening comprehension. This observation is buttressed given the moderate correlations between these two aural subtests and Part 4 (Composition) and the Common and Uncommon Parts variables. The Uncommon Parts variable, like the Common Parts variable, is essentially literacy related since it encompasses the Dictation and the four reading subtests. Only Oral Passage Two and Three are not literacy related. The correlations between the Dictation and Formal Equivalents aural subtests also correlate moderately with the Total score variable. However, this should not appear unusual since the subtests comprising the Common and Uncommon Parts variables are overwhelmingly literacy oriented.

The four aural subtests appear to correlate in one of two ways with the remaining variables in Table 27. The Listening Comprehension and Informal Words subtests correlate relatively weakly with the other test variables, while the Dictation and Formal Equivalents subtests correlate moderately with all the test variables in the table. This pattern renders questionable the construct validity of the listening comprehension part of the test. One would expect each listening comprehension subtest to correlate more uniformly with each variable since the four subtests should be measuring the same construct, listening comprehension.

The three oral passages correlated in a more predictable manner with the battery of test variables than the listening comprehension subtests. Table 28 indicates that Oral Passages One (OPAS1) , Two (OPAS2) and Three (OPAS3) correlated with one another in a fairly moderate way with correlation coefficients ranging between (.3136) and (.5163). Similarly, there is nothing especially striking concerning the correlations between the three oral passages and the four reading subtests. Again the correlations are moderately low ranging from (.1555) to (.3803). A like pattern is also evident between the three oral passages and the four test parts. The oral passages correlate marginally with Part 1 (Aural). In contrast, the correlations between the three oral passages and Part 3 (Reading) and Part 4 (Composition) are slightly less robust than the correlations between the oral passages and Part 1 (Aural). Perhaps the correlation between the aural subtests and oral passages should be higher. On the other hand, the marginal correlation between these two variables is likely due to the weak correlations between the Listening Comprehension and Informal Words subtests and Part 2 (Oral) discussed in the previous table.

Regarding the correlations between the three oral passages and the Common Parts, Uncommon Parts and Total variables, the moderate correlations between the oral passages and the Uncommon Parts variable is probably due to the fact that the Uncommon Parts subsumes Oral Passage Two and Three.

Table 28

Correlations Between Oral Passages & All Remaining Test Variables

	OPAS1	OPAS2	OPAS3
Oral Passage One	1.0000		
Oral Passage Two	0.3136**	1.0000	
Oral Passage Three	0.5163**	0.4608**	1.0000
Orthography: Accents	0.2863**	0.2682**	0.3587**
Orthography: Spelling	0.3803**	0.3156**	0.3919**
Identifying Concepts	0.2594**	0.1194	0.1555
Words in Context	0.3522**	0.2869**	0.4682**
Part 1 (Aural)	0.4681**	0.3519**	0.4465**
Part 2 (Oral)	0.7222**	0.8195**	0.8063**
Part 3 (Reading)	0.3886**	0.3156**	0.4348**
Part 4 (Writing)	0.3162**	0.2659**	0.4300**
Common Parts	0.4895**	0.3288**	0.5098**
Uncommon Parts	0.4765**	0.6347**	0.6605**
Total	0.5291**	0.4883**	0.6235**

With respect to the correlations between the oral passages and the Total score variable, the moderately low correlations are likely due to the nature of the test formats. Recall that the examinee is given two minutes to prepare for passage one, four minutes to prepare for passage two and four minutes to prepare for passage three. Perhaps in preparation for the oral presentations the examinee is identifying and writing down needed key vocabulary. In addition, in the oral part of the exam, the examinee is encouraged to "show off" his or her command of the language as indicated on page 5 of the test booklet. The point is that the Total score variable is skewed towards literacy and a standard variant of the Spanish language which is also called for in the execution of the three oral passages.

Overall, the three oral passages appear to correlate with the other test variables as one might expect. Again, the pattern of correlations is much more predictable than the patterns evidenced in the table above. Consequently, the construct validity of Part 2 (Oral) appears defensible based on the correlation evidence generated thus far.

Following are the correlations between the four reading subtests and the remaining test variables which are summarized in Table 29. In general, the four reading subtests appear to correlate moderately low with one another with the exception of the Orthography: Accents (OACC) and Orthography: Spelling (OSPEL) subtests. The correlation between these two subtests (.7370) coupled with the similar nature of the two subtests suggests that a common linguistic ability is being measured by the two subtests. The four reading subtests also correlate as expected with the four different test parts. The Orthography: Spelling subtest correlates moderately (.7748) with Part 1 (Aural) which is probably due to the spelling requirement inherent to the Dictation and Formal Equivalents aural subtests. The manner in which the four reading measures correlate with Part 2 (Oral) is, as one might expect, somewhat low.

The high positive correlations between the Orthography: Accents (.8752) and Orthography: Spelling (.9224) subtests with Part 3 (Reading) indicate that the performance on either of these two subtests correlate highly with the overall performance of the examinee on the Part 3 (Reading) portion of the test. Consequently, only one of these subtests may actually be needed. Moreover, this pattern is also unfortunate since these two subtests require the least amount of reading on the part of the examinee. Lastly, the four reading subtests appear to correlate moderately with Part 4 (Composition) with correlations ranging between

(.2823) and (.5894). Slightly higher correlations between these two variables would be more desirable since reading and writing share intersecting language abilities.

Table 29

Correlations Between Reading Subtests & Test Parts & Total Score Variable

	OACC	OSPEL	IDCON	WCTEX
Orthography: Accents	1.0000			
Orthography: Spelling	0.7370**	1.0000		
Identifying Concepts	0.3535**	0.4884**	1.0000	
Words in Context	0.4304**	0.5912**	0.5232**	1.0000
Part 1	0.5797**	0.7748**	0.3591**	0.5935**
Part 2	0.3884**	0.4535**	0.2157**	0.4555**
Part 3	0.8752**	0.9224**	0.6247**	0.7224**
Part 4	0.5894**	0.5633**	0.2823**	0.4501**
Common Parts	0.6121**	0.6319**	0.3166**	0.5143**
Uncommon Parts	0.7792**	0.8556**	0.4933**	0.6748**
Total	0.7327**	0.7798**	0.4212**	0.6232**

The correlation patterns between the four reading subtests and the Common Parts, Uncommon Parts and Total score variables are not particularly peculiar. At first sight the correlations between the four reading subtests and the Uncommon Parts variable appear moderate to high, ranging between (.4933) and (.8556). This is probably due in part to the fact that the Uncommon Parts variable subsumes the four reading subtests. The moderate correlations between the Total score variable and the Orthography: Accents and Spelling subtests only lend credence to the observation that the Four Skills Exam is primarily a measure of vocabulary, spelling and grammar.

The four reading subtests appear to correlate more strongly with Part 1 (Aural) than with Part 4 (Composition). Again, this pattern is due to the similarity between the Dictation and Formal Equivalents aural subtests and the Orthography: Accents and Orthography: Spelling reading subtests. There is clearly some redundancy across these four subtests in terms of the constructs being measured, listening comprehension and reading ability. However, the redundancy is more a function of the lack of construct validity of the aural subtests than the reading subtests examined in this section. On the other hand, the construct validity of the Orthography: Accents and Orthography: Spelling is also highly questionable since neither of these two reading subtests require the examinee to read beyond the sentence level.

Table 30 summarizes the correlations between the four parts of the Four Skills Exam as well as the correlations between the four test parts and the Common and Uncommon and Total score test variables. In other words, these correlations provide evidence regarding the performance of the examinees on one part of the exam as compared to their performance on one of the other portions of the exam, including the three created composite test variables.

As stated previously, the correlation between Part 3 (Reading) and Part 1 (Aural) is moderately high, (.7408). Essentially, while these two test parts should be measuring unlike language abilities, they appear to be measuring like language abilities. This again is due to the similarity which exists between the Dictation, Formal Equivalents, Orthography: Accents and Orthography: Spelling subtests. In contrast, none of the three remaining test parts appear to be unusually correlated with any of the remaining test parts.

The moderately high correlation (.7067) between Part 1 (Aural) and the Common Parts variable is likely due to the fact that the latter subsumes three of aural subtests. The relatively high correlation (.9671) between Part 4 (Composition) and the Common Parts

Table 30

Correlations Between Test Parts & Composite Test Variables

	Part 1 (Aural)	Part 2 (Oral)	Part 3 (Rdg)	Part 4 (Comp)
Part 1 (Aural)	1.0000			
Part 2 (Oral)	0.5242**	1.0000		
Part 3 (Rdg)	0.7408**	0.4821**	1.0000	
Part 4 (Comp)	0.5562**	0.4170**	0.6254**	1.0000
Common Parts	0.7067**	0.5432**	0.6766**	0.9671**
Uncommon Parts	0.7858**	0.7784**	0.9005**	0.6406**
Total	0.7945**	0.7014**	0.8278**	0.8973**

variable suggests a redundancy in terms of the language abilities being measured by the two variables. However, Part 4 (Composition) is one of the subtests subsumed under the Common Parts variable. The same is true of the high correlation (.9005) between Part 3 (Reading) and the Uncommon Parts variable. That is, the latter variable subsumes all four of the Part 3 (Reading) subtests.

It is interesting to note the high correlation (.8973) between Part 4 (Composition) and the Total score variable. This is likely due to the nature of the scoring used to score the composition in the last of three steps. Recall that the final set of criteria used to score the composition is grammatical in nature including accentuation, spelling and word choice or vocabulary. With the exception of the Listening Comprehension subtest and the three oral passages, all of the remaining subtests focus on spelling, accentuation and vocabulary. This observation leads to the following inference. In the main, the Four Skills Exam is a discrete point test and not the functional language proficiency test its authors purport it to be. Moreover, given the test's present design, the use of Part 4 (Composition) yields as much information about the language ability of the examinee as the three remaining test parts. The Four Skills Exam may only be a measure of one skill, grammar.

The final table presented which contains correlation data for Form A of the Four Skills Exam is presented in Table 31. Recall that the Common Parts (CMNPRTS) variable consists of those subtests (i.e., Listening Comprehension, Informal Words, Formal Equivalents, Oral Passage One, and the Composition) which are essentially the same across all three forms of the test. The Uncommon Parts (UNCMPTS) variable consists of those subtests (i.e., Dictation, Oral Passage Two, Oral Passage Three, and the four reading subtests) which are different across the three forms of the exam.

Based on the correlations reported in Table 31, two observations are in order. The Common Parts variable and the Uncommon Parts variable are somewhat highly correlated with one another, (.7170). In other words, the performance of the examinee on either composite of subtests is essentially the same. The most salient correlation concerns the relationship between the Common Parts variable and the Total score variable. This correlation (.9509) is obviously quite high. From this one can infer that the examinee need only take the Common Parts subtests since these subtests correlate so highly with the Total

Table 31

Correlations Between Composite Variables

	CMNPRTS	UNCMPTS	TOTAL
Common Parts	1.0000		
Uncommon Parts	0.7170**	1.0000	
Total	0.9509**	0.8976**	1.0000

score. It is redundant to require the examinee to take the subtests subsumed under the Uncommon Parts variable since both variables measure essentially the same skills.

Correlations: Form B

Table 32 contains the first set of correlations for Form B and more specifically for the four aural comprehension subtests and the battery of test variables. Overall, the correlations between the four aural subtests with one another correlated somewhat low. With the exception of the correlation between the Dictation (DICT) and Formal Equivalents (FEQUIV) subtests (.7700), the remaining correlation ranged between (.2432) and (.4325). Again, the moderately high correlation between the Dictation and Formal Equivalents

Table 32

Correlations Between Aural Subtests & Remaining Test Variables

	LCOMP	DICT	IWORDS	FEQUIV
Listening Comprehension	1.0000			
Dictation	0.2502	1.0000		
Informal Words	0.4325**	0.2432	1.0000	
Formal Equivalents	0.3586**	0.7700**	0.2575*	1.0000
Oral Passage One	0.1927	0.3114*	0.0563	0.3072*
Oral Passage Two	0.0556	0.2353	-.0770	0.1718
Oral Passage Three	0.1093	0.3107*	0.0247	0.3020*
Orthography: Accents	0.1985	0.6236**	0.2322	0.6519**
Orthography: Spelling	0.1755	0.8192**	0.2238	0.7133**
Identifying Concepts	0.1697	0.4666**	0.2097	0.3559**
Words in Context	0.3286**	0.6508**	0.2526*	0.7548**
Part 1	0.5881**	0.8840**	0.5197**	0.8738**
Part 2	0.1313	0.3333**	-.0110	0.2969*
Part 3	0.2454	0.8096**	0.2706*	0.7748**
Part 4	0.3322**	0.5089**	0.1626	0.6188**
Common Parts	0.4577**	0.5735**	0.2442	0.6969**
Uncommon Parts	0.2595*	0.7900**	0.1382	0.6966**
Total	0.4145**	0.7185**	0.2211	0.7603**

subtests is probably due to the requirement for the examinee to correctly spell a word in both subtests.

A more noticeable pattern emerges in the set of correlations between the aural subtests and the three oral passages. Both the Listening Comprehension (LCOMP) and Informal Words (IWORDS) subtests correlate quite low with each of the three oral passages. Clearly, this is much more the case for the Informal Words subtests as the correlations range from between -(.0770) and (.0563). Naturally, one would expect the two variables to be more highly correlated since the variables are oral and aural measures.

. There may be different explanations for this problem. First, it may be that this portion of the test is not being administered in a language laboratory as previously stated in Chapter 2. Given this possibility coupled with the multiple-choice format, these two conditions may be making rather difficult for the examinee to perform the test tasks. In addition, some reading is required for the Informal Words aural subtest. Perhaps collectively these sources underlie the low correlations between the Listening Comprehension and Informal Words subtests with the three oral passages. On the other hand, the same sort of pattern emerged in the same correlations for Form A as indicated in Table 27.

With respect to the correlations between the aural subtests and the reading comprehension subtests, it is interesting to note the moderate to high correlations which exist between the Dictation and Formal Equivalents subtests and the four reading subtests. Again, the Dictation aural subtest requires spelling as does the Orthography: Spelling reading subtest. Given the similarity in task type, the correlation between these two subtests (.8192) indicates that the two measures are measuring essentially the same skill.

The Formal Equivalents appeared to correlate moderately high with the Orthography: Spelling (.7133) and Words in Context (.7548) reading subtests. All three subtests focus on the ability of the examinee to spell words or to understand vocabulary. As with Form A, the Dictation and Formal Equivalents subtests appear to be more a measure of written vocabulary than listening comprehension.

The Listening Comprehension and Informal Words aural subtests behave in a like manner as they both correlate low with the four reading subtests. However, these correlations are more in line with what one might anticipate between listening comprehension and reading comprehension subtests.

As one would expect, the four aural subtests correlate moderately to high with their composite Part 1 (Aural). However, and once again, the Dictation and Formal Equivalents subtests correlate quite highly with the composite variable, (.8840) and (.8738), respectively. As noted above, the Listening Comprehension and Informal Words subtests correlated quite low with each of the oral passages and do as well with Part 2 (Oral), the composite oral variable. The -(.0110) correlation indicates that if an examinee did well on the Informal Words subtest, the examinee performed poorly on the oral passages and vice versa. Whichever the direction, this should not be occurring since the Informal Words subtest and the oral passages are intended to measure aspects of a common construct, language proficiency.

The correlations between the four subtests and the Part 3 (Reading) variable are much the same as those discussed above where the correlations were examined on a subtest by subtest basis. That is, the Dictation and Formal Equivalents subtests correlate highly with the reading comprehension measures and the composite reading variable. These same subtests (i.e., Dictation and Formal Equivalents) also correlate moderately with the Part 4 (Composition) variable. Again, this is most likely due to the spelling and vocabulary skills involved in all three measures.

In keeping with this pattern, the Dictation and Formal Equivalents subtests also correlated moderately with the Common Parts, Uncommon Parts and Total score variables. While there should be some shared variance between the aural subtests and these composite variables, the manner in which the Dictation and Formal Equivalents subtests correlated with the Total score variable indicates that the performance of the examinee on these two aural subtests correlates substantially with their overall score. Given that the two aural subtests in question require correct spelling, one may infer that the overall nature of the Four Skills Exam is once again discrete point in nature.

The three oral passages appeared to correlate in a predictable manner with the battery of test variables as reported in Table 33. The oral passages correlated with one another in a fairly consistent manner with the exception of the moderately high correlation (.7368) between Oral Passage One (OPAS1) and Three (OPAS3). Even this seemingly high correlation is acceptable given the fact that each of the three passages follow the same test format and elicit essentially the same linguistic behavior, oral production.

The fairly low correlations between the three oral passages and the four reading subtests should also be expected since oral proficiency is not necessarily positively correlated with reading ability. However, there is some evidence which indicates that the two sets of measures share some common linguistic ground. This is especially clear for the correlations between Oral Passage One and the four reading subtests. To a lesser degree, this also holds true for Oral Passage Three. However, the correlations are somewhat low between Oral Passage Two (OPAS2) and the four reading subtests.

These same patterns emerge when one examines the correlations between the three oral passages and the four test parts. There are relatively low correlations between the oral passages and Part 1 (Aural); high correlations between the oral passages; low correlations with the Part 3 (Reading) variable; and a similarly low set of correlations between the oral passages and the Composition. The correlations for the Common Parts, Uncommon Parts and Total score variables also fluctuate accordingly. The highest set of correlations is found between the oral passages and the Uncommon Parts variable which range between (.6073) and (.6789). These moderate correlations may be due to the fact that

Table 33

Correlations Between Oral Passages & Remaining Test Variables

	OPAS1	OPAS2	OPAS3
Oral Passage One	1.0000		
Oral Passage Two	0.3859**	1.0000	
Oral Passage Three	0.7368**	0.5642**	1.0000
Orthography: Accents	0.3931**	0.2431	0.3417**
Orthography: Spelling	0.3472**	0.1062	0.2179
Identifying Concepts	0.2336	0.0170	0.0871
Words in Context	0.3286*	0.0982	0.2345
Part 1	0.3222*	0.1752	0.2920*
Part 2	0.7913**	0.8330**	0.8873**
Part 3	0.4083**	0.1610	0.2879*
Part 4	0.3247*	0.2489	0.3359**
Common Parts	0.4732**	0.2831*	0.4328**
Uncommon Parts	0.6073**	0.6430**	0.6789**
Total	0.5737**	0.4631**	0.5777**

the Uncommon Parts variable subsumes Oral Passages Two and Three. Overall, the three oral passages appear to correlate predictably with the other test variables.

The correlations between the four reading subtests and the remaining tests variables are reported in Table 34. In general, the four reading subtests correlate in a like fashion with one another with the exception of two moderately high correlations. The correlation coefficient for the Orthography: Accent (OACC) and Orthography: Spelling (OSPEL) subtests (.6732) indicates that these two subtests are measuring a common construct. Again, both subtests deal with related aspects of spelling or orthography. The slightly higher correlation (.7284) between the Words in Context (WCTEX) subtest and the Orthography: Spelling subtest can be interpreted in the same light since the former subtest is more a reading comprehension vocabulary type measure. That is, the Words in Context subtest only requires the examinee to select the correct option which essentially consists of

Table 34

Correlations Between Reading Subtests & Other Test Variables

	OACC	OSPEL	IDCON	WCTEX
Orthography: Accents	1.0000			
Orthography: Spelling	0.6732**	1.0000		
Identifying Concepts	0.4492**	0.5495**	1.0000	
Words in Context	0.4727**	0.7284**	0.4912**	1.0000
Part 1	0.6359**	0.7494**	0.4439**	0.7215**
Part 2	0.3756**	0.2472	0.1176	0.2429
Part 3	0.8369**	0.9404**	0.6659**	0.7875**
Part 4	0.5348**	0.4967**	0.2654*	0.4706**
Common Parts	0.5881**	0.5569**	0.3046*	0.5483**
Uncommon Parts	0.7296**	0.7472**	0.4645**	0.6423**
Total	0.7023**	0.6892**	0.4009**	0.6385**

identifying the correct lexical item to be inserted in a missing blank in a paragraph length text..

With respect to manner in which the four reading subtests correlate with the four test parts, the patterns are much the same as those reviewed for Form A. The four reading subtests correlate from moderate to high with Part 1 (Aural). Again, this is probably due to the Dictation and Formal Equivalents aural subtests which appear to be measuring a linguistic skill closely related to the skill tapped by the reading comprehension subtests. There are relatively low correlations, as one might expect, between the four reading subtests and Part 2 (Oral).

Each of the four reading subtests correlate highly with their composite, Part 3 (Reading). The Orthography: Spelling subtest yielded the highest correlation (.9404) with the composite reading variable. Arguably, this single measure, the Orthography: Spelling subtest, could suffice as the sole subtest for the reading comprehension portion of the test given the discrete point design of this part of the exam and the supporting correlational evidence. With the exception of the Identifying Concepts (IDCON) subtest, the three remaining subtests correlated moderately with Part 4 (Composition). Perhaps the Identifying Concepts subtest correlated somewhat lowly with the Composition since the former subtest, unlike the other three reading subtests, requires more actual reading and transcends the selection of a single word for a response.

The four reading subtests also correlated in a predictable manner with the Common Parts, Uncommon Parts and Total score variables. The four reading subtests correlated more readily with the Uncommon Parts variable since the latter subsumes the four reading subtests. As concerns the correlations between the four reading subtests and the Total Score variable, the Orthography: Accents and Orthography: Spelling subtests correlated somewhat high with the overall performance of the examinee on the Four Skills Exam. As stated previously, this is likely due to the overall discrete point nature of these subtests and the test in general, with the exception of the three oral passages.

Table 35 summarizes the correlations among the four test parts as well as with the Common Parts, Uncommon Parts and Total score variables. The relatively high correlation (.7922) between Part 3 (Reading) and Part 1 (Aural) substantiates the earlier interpretations set forth above regarding the correlations among their respective subtests. These two test parts appear to measuring similar linguistic abilities, primarily spelling and vocabulary.

Table 35

Correlations Between Parts & Composite Test Variables

	PART 1 (Aural)	PART 2 (Oral)	PART 3 (Rdg)	PART 4 (Comp)
Part 1	1.0000			
Part 2	0.3001*	1.0000		
Part 3	0.7922**	0.3202*	1.0000	
Part 4	0.5859**	0.3536**	0.5599**	1.0000
Common Parts	0.6935**	0.4550**	0.6333**	0.9749**
Uncommon Parts	0.7285**	0.7686**	0.8166**	0.5706**
Total	0.7717**	0.6308**	0.7695**	0.8906**

In regard to the near perfect correlation between Part 4 (Composition) and the Common Parts variable, this correlation is likely to be spuriously high since the Common Parts variable encompasses Part 4 (Composition). Similarly, the high correlation between Part 3 (Reading) and the Uncommon Parts variable (.8166) may also be due to the fact that the Uncommon

Parts variable consists, in part, of the four reading subtests. The same holds true for the relatively high correlation (.7686) between Part 2 (Oral) and the Uncommon Parts variable since the latter consists of two of the three oral passages. Some interesting inferences can be made, however, regarding the correlations between the four parts of the exam and the Total score variable. The correlation between Part 4 (Composition) and the Total score variable is high, (.8906). This indicates that if the examinee does well on the composition, he or she does well on the test in general and vice versa. Nonetheless, it is important to bear in mind that the score the examinee receives on the Composition is based essentially on the mechanics of the Spanish language or grammar.

Moreover, if the above inference is correct, this would also help explain the relatively high correlations between both Part 1 and Part 3 with the Total score variable. Part 1 (Aural) correlates readily (.7717) with the Total score variable because both test entities measure essentially the same linguistic trait, spelling and vocabulary. Much the same argument can be made for the high correlation (.7695) between Part 3 (Reading) and the Total Score. This argument is buttressed by two additional observations.

First, recall the high correlation (.7922) between Part 3 (Reading) and Part 1 (Aural). Again, the most plausible explanation for the high correlation between these two test parts is the likelihood that the Dictation, Formal Equivalents, Orthography: Accents and Orthography: Spelling subtests all measure essentially the same construct, spelling. Second, note the correlation (.6308) between Part 2 (Oral) and the Total score variable. This correlation, while somewhat high, correlates least well with the Total score variable as compared to the other three test parts. Moreover, the coefficient of determination for the correlation between Part 2 (Oral) and the Total score is approximately 39% or a little more than one-third of the total variance. The lion's share of the variance is contributed by the remaining three parts of the test which are discrete point in nature and encompassing spelling, vocabulary and grammar like skills.

Given the inferences made regarding the correlations reviewed for both Form A and B of the Four Skills Exam, and the supporting arguments, it appears safe to assume that the construct validity of the test can be called into question. First, it is not the functionally based language test its authors purport it to be. It is more a discrete point grammar test than a language test founded on language functions. While grammar is clearly a requisite for language proficiency, there is no need for three test parts that measure the same linguistic construct, in this case grammatical competence. The prime suspects are Parts 1 (Aural) and Part 3 (Reading). These two test parts are measuring a linguistic skill which more readily belongs under the auspices of Part 4 (Composition). Consequently, the Four Skills Exam appears to be measuring only two Spanish language skills, oral proficiency and composition.

The final correlation table to be examined, Table 36, consists of the correlations between the three created composite variables. The correlation between the Common Parts variable (i.e., Listening Comprehension, Informal Words, Formal Equivalents, Oral Passage One and the Composition) and the Uncommon Parts variable (i.e., Dictation, Oral Passage Two and Three, Orthography: Accents, Orthography: Spelling, Identifying Concepts and Words in Context) is moderate, (.6643). This moderate correlation indicates that the performance of the examinee on the Common Parts (CMNPRTS) variable is similar to their performance on the Uncommon Parts (UNCMPTS) variable. If this correlation had been higher, the implication would be that both variables are measuring a like linguistic ability. Moreover, there would be no need for nearly half of the subtests.

A further inference which can be drawn from Table 36 concerns the correlations between the Common Parts and Uncommon Parts variables to the Total score variable. A high correlation (.9474) exists between the Common Parts variable and the Total score variable. However, this correlation is spuriously high since the Total score variable subsumes the Common Parts variable. On the other hand, the total absence of Part 3 (Reading) from the Common Parts variable does not appear to affect the manner in which the Common Parts variable correlates with the Total score variable. In effect, the Reading Comprehension subtests may be superfluous to the exam given their design. The three aural comprehension subtests must be compensating for the variance contributed by the reading comprehension measures to the Total score since these two sets of subtests have already been shown to correlate moderately high (.7922) with one another.

Similarly, there is a high correlation between the Uncommon Parts variable and the Total score variable, (.8686). However, this correlation is also spuriously high since the Total score variable subsumes the Uncommon Parts variable.

Table 36

Correlations Between Composite Variables & Total

	CMNPRTS	UNCMPTS	TOTAL
Common Parts	1.0000		
Uncommon Parts	0.6643**	1.0000	
Total	0.9474**	0.8686**	1.0000

All of the correlations examined with reference to Form B of the Four Skills Exam generate one critical piece of evidence which is useful in making an overall judgment about the exam. The relatively high correlations which exist between the two aural subtests (i.e., Dictation and Formal Equivalents) and the two reading comprehension subtests (i.e., Orthography: Accents and Orthography: Spelling) indicate that the four subtests are measuring more the same linguistic ability than differing linguistic abilities. This observation is also supported given the moderately high correlation between the two test parts (i.e., Part 1 (Aural) and Part 3 (Reading) of which these subtests form a portion.

Given this correlational evidence, one may infer that the construct validity of these two test parts can be called into question. This author maintains that only the Listening Comprehension subtest measures listening comprehension. The remaining three listening comprehension subtests measure spelling and vocabulary using a tape-mediated format. Similarly, only the Identifying Concepts reading subtest measures reading comprehension. The three remaining subtests measure accentuation, spelling and vocabulary. The unfortunate consequence is that judgments have been made regarding the listening and reading ability of the examinee based on subtests that measure discrete-point skills.

Equivalency of Forms

In order to test for the equivalency of forms of the Four Skills Exam two analyses using the ANOVA statistical procedure were conducted. Recall that two composite variables (i.e., Common Parts and Uncommon Parts) were created. Again, the Common Parts variable consists of a number of subtests which are essentially the same across all three forms of the test; the Uncommon Parts variable consists of subtests which are different across test forms.

For each of the two analyses there was one independent variable, the test form which consisted of two levels, Form A or B. There was also only one dependent variable which consisted of the examinees' mean score on the two composite variables. No significant differences were expected to be found between the examinees' mean scores on either Forms A or B of the test. Significance was tested for at the (.05) level. Lastly, given the small sample of those examinees which took Form C (*n* = 10), only Forms A and B could be tested for equivalency.

No significant differences were found between the performance of the examinees on Forms A and B of the test using the Common Parts dependent variable. The results of this analysis are reported Table 37. There were five missing cases.

Table 37

Common Parts Scores on Forms A & B
(*n* = 202)

	$\bar{x}$	s.d.	n	F
Form A	.501	.216	142	.265
Form B	.550	.241	60	

Table 38

Uncommon Parts Scores on Forms A & B
(*n* = 200)

	$\bar{x}$	s.d.	n	F
Form A	.731	.152	140	.984
Form B	.726	.156	60	

Similarly, significant differences were not found between the performance of the examinees on Forms A and B of the test using the Uncommon Parts dependent variable. The results of this analysis are reported in Table 38. There were seven missing cases.

Summary

The evidence generated through the analyses presented in this chapter do not readily support the construct validity of the aural and reading parts of the Four Skills Exam. The moderate correlations between the aural and reading subtests indicate that a similar linguistic ability (i.e., spelling and vocabulary) is being tapped by these two sets of subtests. This fact makes the interpretation of test scores problematic. That is, the inference an individual makes regarding a passing score on the Aural part of the test is that the examinee can comprehend spoken Spanish. However, a passing score on the Aural part of the exam is a better measure of the examinee's spelling ability and knowledge of vocabulary.

The same case can be made regarding the reading ability of the examinee. The meaning of a passing score on the reading part of the exam seems to be a better indicator of the individual's orthographic skills. The result is, consequently, aural and reading test scores that do not readily reflect the language abilities they purport to. The social consequences are obvious. Examinees passing these parts of the test do not necessarily have the skills the other stake-holders (e.g., students, institutes of higher education, NM State Department of Education, etc.) might believe they have.

CHAPTER 6

ADDITIONAL ASPECTS OF VALIDITY

In this chapter six questions are addressed which generate additional evidence considered essential in setting forth an overall evaluative judgment of the validity of the Four Skills Exam. Some of the evidence sheds further light on the validity of the test, while some of the questions addressed offer insight into the social consequences of the use of the instrument.

ANOVA and MANOVA analyses were conducted using various sociodemographic variables as the independent variables. Where statistical significance was found, post-hoc analyses were also conducted between levels of the variables examined. An alpha level of (.05) was used. The chapter concludes with a brief summary the findings have for the validity of the Four Skills Exam.

Overall Performance of Examinees

The first question addressed did not entail any statistical analyses, but simply asked: How well did the examinees perform on the different parts of the test and the test as a whole?

Recall that the examinee receives either a Pass or Fail score on each part of the exam and must pass all four parts of the test in order to meet the language proficiency criteria established by the New Mexico State Department of Education. Table 39 summarizes the percentages of the examinees passing or failing the different parts of the test and the test as a whole.

Table 39

Pass/Fail Percentages
($n = 217$)

	Part 1 (Aural)	Part 2 (Oral)	Part 3 (Rdg)	Part 4 (Comp)	Total
Fail	64%	21%	54%	67%	80%
Pass	36%	79%	46%	33%	20%

It is important to state that these percentages are based on the final Pass or Fail designation on the Official Score Sheet for each examinee. A small number of the examinees either did not complete the different sub-tests within each section or did not attempt the section at all. These examinees received a fail score under these circumstances.

The data presented in Table 39 indicate that the Four Skills Exam was difficult for the examinees in this data set to pass on the first attempt. As the table reflects, Part 4 (Composition) was the most difficult, and Part 1 (Aural) was only slightly less difficult than the writing measure. Part 3 (Reading) also presents an obstacle for more than half of the examinees. Only Part 2 (Oral) was readily passed by the examinees.

On the whole, the parts of the exam requiring some measure of Spanish language literacy, including the aural part of the test given its spelling and vocabulary orientation, proved to be the most difficult for the examinees.

The Role of Institutes of Higher Education

It is important to consider, albeit briefly, the implications these data have for the Spanish language training the prospective bilingual education teachers receive at their respective institutes of higher education. Clearly, with only a mere 20% of the examinees passing the exam on the first attempt, something is amiss in this language training and testing enterprise, but is it the test, the training or a combination of these and other factors?

Since the institutes of higher education are informed of and given the explicit responsibility of developing the non-English language competencies of prospective bilingual education teachers at their respective institutions, the role the universities and colleges play in this process is key to the validity of the test. From this perspective, if the examinees are not provided with adequate language guidance and training, then the validity of the Four Skills Exam is weakened. A test cannot be a valid measure of an ability for which poor training has been provided.

Table 40 offers data which help answer the question: Did the performance of the examinee vary as a function of institutional affiliation? However, these data must be interpreted with caution since the test site where the examinee took the Four Skills Exam is being used as the institute of higher education where the examinee received his or her Spanish language training.

A MANOVA was conducted using five test sites as the independent variables and the test scores of the examinees on each part of the exam and on the exam as a whole as dependent variables. The results of the MANOVA [*F* (16, 568.88) = 3.02, *p*= .001] indicated that the performance of the examinees varied significantly as a function of test site (i.e., institutional affiliation). Table 40 summarizes the relevant univariate statistics.

Table 40

Univariate Source Table for Site Variable
(*n* = 194)

	Site 1		Site 2		Site 3		Site 4		Site 5		p
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	
Aural	45.49	05.99	40.83	09.19	44.30	07.93	42.45	09.09	46.21	07.76	.018*
Oral	43.35	09.83	40.97	10.37	46.13	10.75	33.31	12.57	45.78	08.19	.001*
Rdg	48.82	07.90	43.23	10.66	46.94	09.84	44.29	09.20	49.23	07.16	.010*
Comp	19.70	23.51	14.00	21.48	22.80	23.44	18.17	23.16	24.00	24.00	.430

*p < .05

The Univariate F-tests indicated that the performance of the examinees varied significantly as a function of test site on Part 1 (Aural), Part 2 (Oral) and Part 3 (Reading) but not on the composition. Post-hoc analyses using Scheffe's formula indicated that Site 1 performed significantly higher than Site 2 and Site 4 on the aural part of the test. Site 3 and Site 5 also scored significantly higher than Site 2 on the aural portion of the test. It seems that the examinees associated with Site 2 were outperformed by those associated with Sites 1, 3 and 5.

Post-hoc analyses for the oral part of the exam indicated that those examinees associated with Site 4 scored significantly lower than those examinees associated with each of the remaining four test sites. The only other significant difference was found between

Site 2 and Site 3. Again, those examinees associated with Site 2 performed significantly lower than those examinees associated with Site 3.

With respect to the post hoc analyses conducted for Part 3 (Reading) of the exam, those examinees associated with Sites 1, 3 and 5 scored significantly higher than those examinees associated with Site 2. Site 1 and Site 5 also scored significantly higher than Site 4 on this part of the test.

An ANOVA was also performed which indicated that there was a significant difference (f= .019) between the Total score of the examinees as a function of the Test Site variable. Post-hoc analyses revealed that Sites 1, 3 and 5 scored significantly higher than Sites 2 and 4.

From these findings one can at least infer that the institutions of higher education are not giving the examinees uniform opportunities to develop the language abilities measured by the Four Skills Exam, especially at Sites 2 and 4. On the other hand, one must temper this finding with the fact that only twenty percent of the examinees in this data set passed all four parts of the exam on the first attempt. In addition, none of the various institutions of higher education distinguished themselves in preparing the examinees for the written portion of the exam, the composition.

Formal Spanish language training

The third question investigated concerns the influence formal Spanish language training might have on the performance of the examinees on the Four Skills Exam. One would expect that the those examinees with more formal Spanish language training would perform better on the Four Skills Exam than those with less or no such training.

In the following MANOVA, the independent variable was formal Spanish language training which consisted of three levels: (Group 1) no formal language training, (Group 2) formal language training in either high school or college, and (Group 3) formal language training in high school and college. Again, the dependent variables were the mean scores on each of the four parts of the test.

The MANOVA indicated that there was a significant difference [$F(8, 408) = 3.64$, $p = < .001$] in the performance of the three groups as a function of formal language training. Table 41 contains relevant univariate statistics.

Table 41

Univariate Source Table for Formal Study of Spanish

	No Study (*n* = 14)		HS or College (*n* = 79)		HS & College (*n* = 117)		
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	p
Aural	41.42	07.54	41.44	07.66	45.27	08.09	.003*
Oral	44.85	11.24	37.73	11.87	44.09	10.01	.001*
Rdg	42.31	12.11	44.09	10.00	48.21	08.38	.003*
Comp	15.42	25.09	16.25	21.31	20.00	23.47	.471

*$p < .05$

As Table 41 reveals significant differences were found for the first three parts of the exam but not on the composition. Post-hoc analyses for the aural part of the exam indicated that those examinees in Group 3 performed significantly higher than those examinees in Group 2. In the comparison of the mean scores on the aural part of the test between Group 1 and Group 3, significance was not reached by a mere difference of (.01) between t critical (1.74) and t observed (1.73).

On the oral part of the exam, Group 1 and Group 3 both scored significantly higher than Group 2. As for the reading portion of the exam, Group 3 performed significantly higher than both Groups 1 and 2.

An ANOVA was conducted to determine whether or not the Total Score varied across the three levels of formal study of Spanish. Significance was found beyond the (.01) level. Consequent post-hoc analyses indicated that those examinees in Group 3 performed significantly higher on the test as a whole than those examinees in Group 2 but not Group 1.

In short, Group 3, consisting of those examinees which reported having studied Spanish formally in high school and college, performed significantly higher than Group 2 on the aural, oral and reading parts of the test and on the test as a whole. In contrast, Group 3 only performed significantly better than Group 1 on the reading portion of the test. Lastly, Group 1 scored significantly higher than Group 2 on the oral part of the test.

The performance of Group 1, the group with no formal study of Spanish, on the oral part of the exam merits explanation. Perhaps this group consists of native speakers of Spanish who feel they have a fair command of the oral language but for different reasons did not study Spanish in an educational setting.

There seems to be some evidence that the more formal study of Spanish a prospective bilingual education teacher has, the better this person will do on the Four Skills Exam. On the other hand, it is not known just how much formal study is needed. Nonetheless, it seems safe to assume that the formal language training the examinees in this data set received was not enough given their overall poor performance on the test discussed at the beginning of this chapter.

Spanish Language Background

The fourth question concerns the performance of the examinees on the Four Skills Exam with varying Spanish language backgrounds. That is, did those examinees with a native Spanish language background perform better than those without such a background?

The independent variable, Spanish language background, consisted of three groups: (Group 1) examinees reporting not speaking Spanish as they grew up or presently at home, (Group 2) examinees reporting speaking Spanish presently at home but not as they grew up, and (Group 3) those examinees that reported speaking Spanish as they grew up and presently at home. The dependent variables were the mean scores of the examinees on the different parts of the test.

Again, the MANOVA indicated that there was a significant difference in the performance of the three groups on the different parts of the exam [F (8, 408) = 4.67, $p <$ .001]. The standard deviations, means and p values for the univariate analyses are summarized in Table 42. The univariate analyses indicated that the groups differed in performance on only the reading comprehension part of the test. Interestingly, post-hoc analyses indicated that those examinees in Group 1 and Group 2 scored significantly higher than those examinees in Group 3 on the reading portion of the test. It should also be noted that the ANOVA conducted indicated that there was not a significant difference among the three groups on the Total Score variable.

Table 42

Univariate Source Table for Spanish Language Background

	Group 1 (*n* = 47)		Group 2 (*n* = 49)		Group 3 (*n* = 114)		
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	p
Aural	42.63	09.32	44.38	08.01	43.61	07.60	.571
Oral	40.91	09.39	42.16	11.21	41.92	11.96	.839
Rdg	49.32	09.42	47.38	09.09	44.54	09.41	.009*
Comp	20.68	24.85	15.18	22.28	18.63	22.12	.485

*$p < .05$

Ironically, when all of these analyses are taken together, they seem to indicate that those examinees who did not speak Spanish as they grew up nor speak Spanish at home when they took the test scored significantly higher on the reading part of the test than those examinees who did speak Spanish. In effect, there does not appear to be any notable advantage for those examinees with a Spanish language background. On the other hand, it should be kept in mind that the test as a whole does not always measure up to the highest standards of validity.

Geographic Location

The performance of the examinees on the test was also examined using the geographic location in which the examinees reported residing as the independent variable. Did the performance of the examinees vary as a function of the region of New Mexico in which they happened to reside?

The residence variable consisted of three levels: north, central and south. The MANOVA indicated that there was a significant difference in the performance of the examinees as a function of the residence variable, [F (8.00, 384.00) = 2.25, $p <$ .023]. Relevant univariate statistics are summarized in Table 43. The data in this table indicate that the performance of the examinees varied significantly as a function of the residence variable on the aural, reading and composition parts of the test. With respect to the aural part of the exam, post-hoc analyses indicated that those examinees residing in the Central and Southern region performed significantly better on the aural part of the exam than those examinees residing in the Northern region.

Table 43

Univariate Source Table for Residence Variable

	North (n = 73)		Central (n = 73)		South (n = 52)		
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	p
Aural	41.42	08.58	45.16	07.25	43.81	08.16	.019*
Oral	40.51	09.78	41.94	11.12	42.67	12.28	.525
Rdg	43.55	10.44	48.85	07.65	45.98	09.97	.003*
Comp	11.34	19.38	21.86	24.28	22.38	22.92	.005*

*$p <$ .05

The examinees associated with the Central region also scored significantly higher than those examinees from the Northern region on the reading component of the test. The Central and Southern region also scored significantly higher than those examinees from the Northern region on the composition part of the test.

The same pattern held for the performance on the test as a whole. The ANOVA using the residence factor as the independent variable and the total score as the dependent variable was significant (f=.005). Consequent post-hoc analyses indicated that the Central and Southern regions performed significantly higher than the Northern region on the test as a whole.

The most salient inference that can be drawn from these analyses is that those examinees reporting residing in the northern region of the state of New Mexico are not performing on par with the examinees from the remaining two regions, with the exception of the oral part of the exam. Perhaps the examinees associated with the northern region also received their Spanish language training at one of the institutes of higher education found to be offering less than desirable training.

Ethnicity and Test Performance

The final set of analyses conducted in this dissertation concerns the relationship between the ethnicity of the examinees with test performance. The final question is: Did the performance of the examinee vary as a function of the Hispanic or non-Hispanic ethnicity of the examinees? Recall that the ethnicity of the examinees was inferred by the surname of the test-takers and each examinee was categorized as either Hispanic or non-Hispanic.

Table 44 contains the results of five ANOVA tests using ethnicity (i.e., Hispanic or non-Hispanic) as the independent variable and the four test parts as well as the total score as dependent variables. As the data indicate, significant differences were found between the two groups on the Reading, Composition and Total Score variables as a function of ethnicity.

Table 44

ANOVA Source Table for Ethnicity Variable

	Non-Hispanic		Hispanic			
	$\bar{x}$	sd	$\bar{x}$	sd	F	P
	(*n* = 56)		(*n* = 159)			
Aural	44.12	09.01	43.03	08.08	0.705	.401
	(*n* = 54)		(*n* = 160)			
Oral	43.35	11.08	40.89	11.66	1.837	.176
	(*n* = 56)		(*n* = 158)			
Rdg	49.43	09.97	45.00	09.28	9.040	.003*
	(*n* = 56)		(*n* = 160)			
Comp	27.21	24.48	15.15	21.30	12.291	.001*
	(*n* = 54)		(*n* = 156)			
Total	68.55	18.92	60.34	16.45	9.215	.002*

*$p < .05$

Again, it seems counter-intuitive that those examinees judged to have an Hispanic surname essentially fared less well on the Four Skills Exam than those examinees without an Hispanic surname. As with the Spanish language background variable, there was no apparent advantage for Hispanic surnamed examinees as concerns their performance on the test and more specifically on the aural and oral parts of the test.

Summary

The purpose of these analyses clearly have implications for the validity of the exam and shed some light on the social consequences of using the Four Skills Exam in New Mexico. The social consequences of the relatively high fail rate on the exam as a whole is clear. The need for bilingual education teachers in New Mexico is exacerbated and consequently the needs of those school aged students who could benefit from bilingual instruction go unmet.

The pass/fail data also indicate that bilingual education teachers are not well prepared by institutes of higher education to meet the Spanish literacy demands in particular as measured by the Four Skills Exam. As stated previously, a test cannot be valid if the individuals being tested have not been given the opportunity to develop the abilities being tested. This aspect of validity, however, is external to the instrument and not a property of the test itself.

To complicate matters, the Spanish language training offered to prospective bilingual education teachers in New Mexico appears to vary somewhat from institution to institution. This is troublesome since the Spanish language competencies Spanish language training should address are the same for all institutions in the state of New Mexico. These data suggest that the manner in which the institutes of higher education are attempting to meet these competencies varies in quality at least among some of the institutions. The social consequences are an unequal opportunity for the examinees to develop the desired Spanish language skills. Again, this is an aspect of validity external to the test itself.

The relationship between formal Spanish language training and test performance should send a clear signal to prospective bilingual education teachers and institutes of higher education. The examinees in this data set could have clearly benefited from more Spanish language training, especially in the Spanish literacy skills area. This especially holds true for those stake-holders in the northern region of the state.

On the other hand, there is a certain paradox which emerges when one considers the analyses related to the Spanish language background of the examinees. One would expect the examinees with more experience with the Spanish language to do better on a Spanish language proficiency measure than those examinees without such a background. Here the construct validity of the Four Skills Exam becomes suspect.

All evidence up to this point appears to support that the Four Skills Exam is a discrete-point Spanish language proficiency measure. Consequently, it is not surprising that those examinees with more formal Spanish language training performed better than those with less such training. The training and test are more closely aligned with one another than are the Spanish language experiences and knowledge native speakers bring with them to the task. The point is, language proficiency is more than only grammatical competence and this competence is basically what the Four Skills Exam measures.

The social consequences of making judgments about the Spanish language proficiency of prospective bilingual education teachers based primarily on their grammatical competence are straight-forward. Bilingual education teachers must also be able to demonstrate social (e.g., non-verbal communication) and functional competence (e.g., ability to deliver instruction) in the Spanish language. In brief, it may well be that the examinees who reportedly grew up speaking Spanish and still speak Spanish at home have social and functional abilities in the language that are not measured by the Four Skills Exam.

The analyses related to the ethnicity of the examinees also generate social consequences of the Spanish language proficiency testing enterprise under consideration. It is ironic that the prospective bilingual education teachers who are most likely to share linguistic, cultural and socio-historical backgrounds with the student population experienced the most difficulty in passing this exam. Consequently, the source of potential bilingual education teacher role models for the targeted student population is impeded.

CHAPTER 7

CONCLUSIONS AND FUTURE RESEARCH

This dissertation investigated the unified validity of the Four Skills Exam (Messick, 1989). A variety of evidence was generated which revealed certain weaknesses in the test and its applications. Some of the evidence is internal to the test itself and stems from the psychometric properties of the instrument. Additional evidence, however, is external to the test and linked to how the test is actually managed and used in practice. The evidence is presented below following the framework advanced by Messick. This chapter concludes with a general summary statement and directions for future research.

I. Construct validity

The construct validity and behavioral relevance of the Four Skills Exam, based on this research, has not in fact measured up to the level required of such a high-stakes test. Nor has it met the standards laid down initially by its proponents and designers. While it was the intent of the original test design team to develop a test which captured the real life language demands of the bilingual classroom, it is difficult to see how those demands are actually reflected (if at all) in the manner in which language is processed as the examinee works through the Four Skills Exam. The tasks required by the test hardly conform to the sorts of theories of language proficiency advocated by Bachman, Oller, and others. In brief, the test falls short of operationalizing the constructs (i.e., aural, oral, reading and writing proficiency) the designers set out initially to measure. As regrettable as it may be, the Four Skills Exam is more appropriately characterized as a test of grammar, spelling, and vocabulary, even tending toward the discrete-point end of the spectrum, rather than one reflecting real life demands or ordinary classroom discourse.

In general the reliabilities observed in this study fell short of the desired .90 mark recommended for such a high-stakes test. The objectively scored parts of the test (i.e., the aural and reading parts) yielded less than acceptable reliability coefficients; the reliability of the subjectively scored parts of the test (i.e., the oral and written parts), on the other hand, for reasons given in Chapter 4, were probably spuriously high owing to a halo effect. The fault there appeared to be poorly explicated scoring criteria and procedures. Taking into account that a test that is not reliable cannot be valid (Bachman, 1990), the subjectively scored parts of the Four Skills Exam become doubly suspect.

The evidence related to the analyses which produced the correlation coefficients also indicate that the aural and reading parts of this test correlate too readily with one another. Again, this is probably due to the fact that the subtests for these two parts of the test focus on spelling, vocabulary and orthography (discrete language skills) which are not directly related to the constructs targeted for measurement (i.e., listening comprehension and reading ability).

The analyses conducted to examine the relationship between formal language training and the test scores of the examinee also lend support to the discrete-point nature of the test. If the formal language training the examinees received can be characterized as traditional with a focus on the mastery of the formal structural aspects of the Spanish language, then it should come as no surprise that those individuals with more formal training performed significantly better on the discrete point portions of the test.

Similarly, the analyses related to the native language background variable reveal that even those examinees who reportedly speak Spanish presently and spoke Spanish as they grew up performed no better on the aural or oral parts of the test than those examinees lacking these attributes. Again, if the Four Skills Exam were real life oriented, then one would expect that those with the most real life experience with the Spanish language would score consistently higher on at least the aural and oral measures.

The final observation to be made regarding the construct validity of the Four Skills Exam concerns the lack of scoring rubrics and benchmarks for the oral and composition portions of the test. The lack of specificity of linguistic criteria underlying the rating of the oral and written discourse generated by the examinee must also detract from the construct validity of these two parts of the test. Based on the foregoing evidence, the construct validity and behavioral relevance of each part of the Four Skills Exam leaves room for improvement.

II. Content relevance and coverage

While the original intent of the test development team was to embed the Spanish language functions the team initially identified into the test, much was apparently lost in the transition. The test seems to focus on spelling, vocabulary, and grammar. This fact severely limits its content relevance and the achieved coverage of what the test was intended to measure.

Only the content of the Listening Comprehension subtest is relevant to the aural content originally targeted by the test development team. In the case of the oral part of the test, no content is entailed which might give a more direct indication of the ability of the examinee to deliver instruction across the curriculum. Similarly, the content of the reading portion of the exam is given over primarily to orthographic aspects of the Spanish language not reading comprehension. However, in terms of content, the composition portion of the exam does appear to cover the original content targeted by the test development team.

With respect to the content inherent to the native language competencies adopted by the New Mexico State Department of Education which became effective in 1989, the Four Skills Exam was never intended to uphold the content associated with these competencies. In short, there is no formal measure in place to ensure that prospective bilingual education teachers can demonstrate the competencies endorsed by the state. From this vantage point, it is not clear why the Four Skills Exam is still being used in its original form. Based on the results reported here, it would appear that the potential usefulness of the test was further distorted when it was adopted to assess the competencies endorsed by the state. In fact, the research shows that the test was actually short-circuited from the beginning when it was put into practice.

III. Value implications of score interpretation

Given the above discrepancies, and based on the original intent of the Four Skills Exam, it still proves difficult to make *valid* inferences regarding the language abilities of the examinees based on their test scores. Some of the examinees may in fact have adequate levels of language proficiency but may not be able to 'pass' the different parts of the test. Similarly, some of the examinees who do pass the test may not have the skills the various stake-holders believe the examinees possess. In short, the meaning of the test scores is blurred due to its psychometric shortcomings.

Related to this issue of score interpretation is the confusion surrounding the grade levels for which the test scores might be valid. Presently, the test is used to endorse bilingual education teachers K-12, a purpose for which the test was never intended. Valdés (1989) provides ample evidence that the Four Skills Exam was intended for bilingual education teachers teaching young children. If the integrity of the test scores is questionable for bilingual education teachers teaching young children, what little validity the test scores do have is fallacious for those teachers delivering instruction at the higher grade levels. Equally critical, the meanings of the scores generated by the Four Skills Exam are not compatible with the native language competencies adopted by the state approximately five years ago.

IV. Social consequences for applied decision-making

The social consequences of using the pass and fail scores generated by the examinees taking the Four Skills Exam for endorsement purposes are undesirable. The intent driving the development and adoption of the Four Skills Exam was to protect the rights of Spanish-speaking children in need of instruction through the medium of their native language. Given the evidence set forth, there is little reason to believe that the Four Skills Exam is adequately fulfilling this intended social function.

Consequently, the educational needs of the school-aged New Mexican community are not well served by the exam in its present form. If this is the case, then the sector of the New Mexican community the Four Skills Exam was intended to protect may be suffering the most from use of this test. The general community which eventually absorbs the students as they graduate from or leave school also endures undesirable consequences. The students may not have received the quality of educational instruction needed in order to contribute their full potential to New Mexico and the broader society. To agitate the situation, there is some evidence which indicates that those examinees who are Hispanic surnamed, grew up speaking Spanish, still speak Spanish at home, and are native to New Mexico experience difficulty in passing the Four Skills Exam. Given the balance of evidence which does not support the unified validity of the Four Skills Exam, the fact that prospective bilingual education teachers with the above characteristics are not readily being endorsed due to their perceived Spanish language proficiency is disturbing. These individuals represent the pool of potential role models for the students in need of Spanish language instruction. Moreover, these individuals - providing appropriate Spanish language training has been rendered - are likely to achieve the highest levels of Spanish language proficiency since they come to the task with a foundation on which to build.

Summary

It is important to highlight one critical point related to the unified validity of the Four Skills Exam. The problems with the test transcend its psychometric properties. Those individuals who have been responsible for making policy decisions regarding its use have not effectively met their responsibility. The test has been used for two purposes for which it was never intended—(1) it has been applied at secondary levels when it was intended for application at the elementary level, and (2) it has been pressed into service to uphold native language competencies adopted years after the test was developed and with which it has little or nothing in common. In addition, policy makers have operated under the assumption that the Four Skills Exam could maintain its integrity over time in the absence of routine test maintenance.

Lastly, those institutions of higher education which train prospective bilingual education teachers must accept their share of the responsibility. Recall, these institutions have the responsibility of moving prospective bilingual education teachers toward explicit native language competencies which reflect a fairly high standard of language proficiency. Nonetheless, in the majority of the cases, the examinees in this data set could not manage a test designed for teachers of young learners. It is important to bear in mind that a test can hardly be valid for learners who have not been given an opportunity to develop the abilities being measured. Institutions of higher education are not only not providing the prospective bilingual teachers with the needed training, but the Four Skills Exam as it is presently used, impedes progress in that direction.

In closing, it ought to be kept in mind that neighboring states (e.g., Utah, Kansas, Oklahoma, Nevada, Oregon, etc.) have scarcely begun to consider the need for qualifying bilingual teachers in terms of their non-English language proficiency. The Four Skills Exam, even with its shortcomings, represents a noble effort on the part of the test development team to protect the rights of students to a meaningful education. Hopefully, between the efforts of the test development team and the analyses conducted in this dissertation, New Mexico can move forward in this arena.

An important step which represents an opportunity to rectify the present situation in New Mexico has recently been taken. In 1994 the New Mexico State Legislature passed House Bill 224 which created and financially supports a position for a full-time bilingual assessment professional whose primary responsibility will be to revise the Four Skills Exam. Hopefully, this individual will also be able to create an open line of communication among the policy-making entities and the institutions of higher education providing Spanish language training. The unified validity of the forthcoming revised Four Skills Exam is contingent upon all responsible parties working in tandem with one another without losing sight of the purpose of the test.

Future Research

With specific reference to this study, there is a need to understand the qualitative dimensions of the development of the Four Skills Exam. As previously stated in the Limitations of the Study, this dissertation did not entail the use of questionnaires, interviews, observations, and other techniques that would have generated a wealth of qualitative information which clearly would have aided in setting forth a more exhaustive analysis of the evidence underlying the unified validity of the instrument in question. Valdés (1989) makes reference to the political controversy in New Mexico which was caused by raising the issue of the prospective bilingual education teacher's Spanish language proficiency and its measurement. It would prove insightful to better understand the political power struggles which occurred during the period that the test was being developed.

In a more general sense, there are numerous and basic areas of research which need to be explored in order to advance the art and science of the measurement of non-English language proficiency of prospective bilingual education teachers. The most immediate need is in the following areas:

1. It is imperative that more research be conducted to investigate the relationship between the Spanish language proficiency of the bilingual education teacher and student achievement. To this author's knowledge, only one research effort (Merino, Politzer, and Ramirez, 1979) has carefully examined this relationship.
2. It is imperative that more research be conducted to investigate the uses to

which the Spanish language is put in a bilingual education setting, especially at the secondary level. Of the Spanish language proficiency exams reviewed in this study and in use in the Southwest, none of the tests report conducting field observations at the secondary level. Perhaps, through such research, it would be possible to better understand the present vagueness surrounding what constitutes an "eighth grade" level of proficiency.

3. It is imperative that more research be conducted to investigate the relationship between the Spanish language training prospective bilingual education teachers receive and their performance on language proficiency measures. It is critical to the field of bilingual education teacher training to begin to identify "best practices" for developing the prospective bilingual education teacher's non-English language proficiency. The research conducted by Milk (1991) is insightful in this respect.

4. It is imperative that more research be conducted to investigate the concurrent validity of Spanish language proficiency tests being used for similar purposes across the country. To this author's knowledge, no such research has been conducted.

5. It is imperative that more research be conducted to investigate the effect the sociolinguistic milieu and language policies have on the Spanish language proficiency of prospective bilingual education teachers of Hispanic descent.

6. It is imperative that more research be conducted which is related to the

development of "authentic" language proficiency measures. Again, the tests presently used in the Southwest for the purpose in question are all paper and pencil tests. There is a need to begin exploring what more direct performance based measures might look like.

In closing, it is incumbent upon bilingual educators to ask simple yet difficult questions such as the one addressed in this study. Moreover, bilingual educators must begin to set and strive towards high standards whatever their role may be. It is my belief that we have yet to discover the full power of bilingual education in the United States. Too much is at stake not to pursue this full research agenda with anything less than full vigor and commitment.

References

Ada, A. F. (1986). Creative education for bilingual teachers. Harvard Educational Review, 56, 386-394.

August, D. & García, E. (1988). Language minority education in the United States: Research, policy and practice. Springfield, IL: Charles C. Thomas.

Bachman, L. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

Barkin-Riegelhaupt, F. (1985). Testing bilingual language proficiency: An applied approach. In L. Elías-Olivares (Ed.), Spanish language use and public life in the USA (pp. 165-180). Mouton Publishers.

Benjamin, R. & Navarrete, C. (1992). Response to the failure rate of New Mexico's bilingual education teachers on the Four Skills Exam. Unpublished manuscript, New Mexico Highlands University, Evaluation Assistance Center-West, NM.

Bills, G. (1988). The US Census of 1980 and Spanish in the Southwest. International Journal on the Sociology of Language, 79. 11-28.

Borg, W.R. & Gall, M.D. (1979). Educational research (3rd ed.). NY: Longman.

California Commission on Teacher Credentialing. (1992) . BCC Bilingual certificate of competence examination (Spanish). Examination Bulletin. Sacramento, CA.

Cárdenas, M. (1981). [Evaluation of "Spanish language proficiency exam"]. Unpublished data. New Mexico State University, NM.

Carrasquillo, A.L. & Segan, F. (1982). Balancing English/Spanish language skills proficiency in bilingual teacher trainees: Some practical approaches. (ERIC Document Reproduction Service No. ED 246 668).

Center for Applied Linguistics. (1974). Guidelines for the preparation and certification of teachers of bilingual/bicultural education. U.S. Office of Education, Washington, D.C.

Clark, E.R. (1990). The state of the art in research on teacher training models with special reference to bilingual education teachers. In Proceedings of the First Research Symposium on Limited English Proficient Student's Issues (pp. 361-391). Office of Bilingual Education and Minority Language Affairs, Washington, D.C.

Cohen, A. (1994). Assessing language ability in the classroom. Boston, MA: Heinle & Heinle.

Crawford, J. (1989). Bilingual education: History, politics, theory, and practice. Trenton, NJ: Crane.

Cummins, J. (1989). Empowering minority students. California Association for Bilingual Education.

Davies, A. (1990). Principles of language testing. Cambridge, MA: Basil Blackwell Ltd.

Gaarder, B.A. (1977). Teacher training for Spanish-medium work in United States schools. In A.B. Gaarder (Ed.), Bilingual schooling and the survival of Spanish in the United States (pp. 81-94). Rowley, MA: Newbury House.

García, E. (1992). Teachers for language minority students: Evaluating professional standards. In Proceedings of the Second National Research Symposium on Limited English Proficient Student Issues: Focus on evaluation and measurement. Office of Bilingual Education and Minority Language Affairs, Washington, D.C.

Garza, A.S. & Barnes, C.P. (1989). Competencies for bilingual multicultural teachers. The Journal of Educational Issues of Language Minority Students, 5, (Fall), 1-25.

Grenier, G. (1984). The effects of language characteristics on the wages of Hispanic-American males. The Journal of Human Resources, XIX, (1), 36-52.

Hannum, T. (1993a). Four Skills Exam Cover Sheet. Unpublished manuscript.

Hannum, T. (1993b). The Four Skills Exam. Unpublished manuscript.

Hernández-Chávez, E. (1988). Language policy and language rights in the United States: Issues in bilingualism. In T. Skutnabb-Kangas & J. Cummins (Eds.), Minority education: From shame to struggle (pp. 45-56). Philadelphia, PA: Multilingual Matters.

Hernández-Chávez, E. (1993). Native language loss and its implications for revitalization of Spanish in Chicano communities. In B.J. Merino, H.T. Trueba, & F.A. Samaniego (Eds.), Language and culture in learning: Teaching Spanish to native speakers of Spanish (pp. 58-74). Washington, D.C.: Falmer.

Hughes, A. (1989). Testing for language teachers. New York, NY: Cambridge University Press.

Kjolseth, R. (1983). Cultural politics of bilingualism. Transaction Social Science and Modern Society, 20, 4.

Lyons, J. (1990). The past and future directions of federal bilingual education policy. ANNALS, The Annals of the American Academy of Political and Social Sciences, 508, (March), 66-80.

Marshall, D.F. & Gonzalez, R.D. (1988). Una lingua, una patria?: Is monolingualism beneficial or harmful to a nation's unity? In K. Adams & D.T. Brink (Eds.), Perspectives on official English: The campaign for English as the official language of the USA (pp. 29-51). Berlin: Mouton de Gruyter.

McFerren, M. (1988). Certification of language educators in the United States. (Educational Report 11). Center for Language Education and Research, Los Angeles, CA. (ERIC Document Reproduction Service No. ED 291 244).

Merino, B., Politzer, R., & Ramirez, A., (1979). The relationship of teacher's Spanish proficiency to pupil's achievement. NABE Journal, 31, (2), 21-33.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18, (2), 5-11.

Milk, R.D. (1991). Preparing teachers for effective bilingual instruction. In M. McGroarty & C. Faltis (Eds.), Languages in school and society (pp. 267-280). Berlin: Mouton de Gruyter.

National Association for Bilingual Education. (1992). Professional standards for the preparation of bilingual/multicultural teachers. Washington, D.C..

National Commission on Testing and Public Policy. (1990). From gatekeeper to gateway: Transforming testing in America. Boston College, Chestnut Hill, MA.

New Mexico State Board of Education. (1987). Competencies for teachers of bilingual education. New Mexico Department of Education, Santa Fe, NM.

Oller, J.W., Jr. (1991). Language and bilingualism: More tests of tests. Bucknell University Press. Associated University Presses.

Oller, J.W., Jr. & Damico, J.S. (1991). Theoretical considerations in the assessment of LEP students. In E.V. Hamayan & J.S. Damico, (Eds.), Limiting bias in the assessment of bilingual students (pp. 77-110). Austin, TX: Pro-ed.

Peñalosa, F. (1980). Chicano bilingualism and the world system. In R.V. Padilla (Ed.), Ethnoperspectives in bilingual education research, Vol. II: Theory in bilingual education (pp. 3-17). Ypsilanti, MI: Eastern Michigan University.

Phillipson, R. (1988). Linguicism: Structures and ideologies in linguistic imperialism. In T. Skutnabb-Kangas & J. Cummins (Eds.), Minority education: From shame to struggle (pp. 339-358). Philadelphia, PA: Multilingual Matters.

Piatt, B. (1990). ¿Only English? Albuquerque, NM: University of New Mexico Press.

Popham, W.J. (1990). Modern educational measurement: A practitioner's perspective. Englewood Cliffs, NJ: Prentice Hall.

Rorro, C.M. (1981). Oral language proficiency assessment for bilingual and English as a second language certification in New Jersey. Trenton, NJ: New Jersey State Department of Education. (ERIC Document Reproduction Service No. ED 260 583).

Ruíz, R. (1988). Bilingualism and bilingual education in the United States. In C.B. Paulston (Ed.), International handbook of bilingualism and bilingual education (pp. 539-560). Greenwood Press.

Shohamy, E. & Reves, T. (1985). Authentic language tests: Where from and where to? Language Testing, 2, 48-59.

Solé, Y. (1990). Bilingualism: Stable or transitional? The case of Spanish in the United States. International Journal of the Sociology of Language, 84, 35-80.

Spener, D. (1988). Transitional bilingual education and the socialization of immigrants. Harvard Educational Review, 58, (2), 133-153.

Spolsky, B. (1985). The limits of authenticity in language testing. Language Testing, 2, (1) 31-39.

SPSS-X User's Guide (3rd edition). (1988). SPSS, Inc.

Stansfield, C. & Kenyon, D.M. (1991). Development of the Texas Oral Proficiency Test (TOPT). (Final Report). Washington, D.C.: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 332 522)

Texas Education Agency. (1988). Study guide Bilingual education 15: Examination for the certification of educators in Texas (ExCET). Austin, TX.

Tienda, M. & Neidert, L. (1984). Language, education, and the socioeconomic achievement of Hispanic origin men. Social Science Quarterly, 65, 519-536.

Trueba, H.T. (1989). Raising silent voices: Educating linguistic minorities for the 21st century. Boston, MA: Heinle & Heinle.

Valdés, G. (1989). Testing bilingual proficiency for specialized occupations: Issues and implications. In B.R. Gifford, (Ed.), Test policy and test performance: Education, language and culture (pp. 207-229). Norwell, MA: Kluwer Academic Publishers.

Veltman, C. (1988). The future of the Spanish language in the U.S. Hispanic Policy Development Project, New York City & Washington, D.C.

Young, R., Minnick, K.F., & Gregory, C. (1986). The Four Skills Exam (FSE): A measure of Spanish proficiency for the classroom. Unpublished manuscript, University of New Mexico, Testing Division, NM.